Facultad
de
Ciencias

NORMALIDAD DE LA ALTURA DE LAS OLAS DEL MAR

Normality of the sea waves

Trabajo de fin de Grado
para acceder al

GRADO EN MATEMÁTICAS

Autora: Marta Ferrero Díez

Directora: Alicia Nieto Reyes

Junio 2020

Quisiera agradecer en primer lugar a la profesora Alicia Nieto Reyes, por haber sido mi directora y guía durante el largo proceso que ha sido este trabajo de fin de grado.

Gracias a mi familia por el día a día, por haberme entregado su apoyo constante y haber confiado en mí a lo largo de este último camino. Sin ellos no habría salido adelante.

Por último, pero no menos importante, no puedo dejar de mencionar a todos mis amigos que han estado a mi lado aportándome fuerzas en esta larguísima etapa que parecía no llegar a su fin. Seguramente la mayor parte no se habrá dado cuenta pero, cada uno de ellos, me ha estado dando empujoncitos constantes para llegar a la meta. Siempre estaré en deuda.

Índice

Resumen

El presente proyecto tiene como finalidad estudiar si las alturas de las olas del mar siguen un proceso Gaussiano. Se va a realizar el estudio de las observaciones almacenadas en 11 boyas distintas de alrededor del mundo. Para realizarlo, se han seleccionado de cada boya 2305 observaciones distintas consecutivas comprendidas entre las 14h y 15h del día 23 de Julio de 2018. Es conveniente conocer la distribución de este fenómeno ya que muchos de los elementos que se encuentran en el mar como son las infraestructuras de petróleo, espigones o incluso los barcos, requieren un modelaje previo a su construcción para determinar, por ejemplo, la resistencia de los mismos. Para poder llegar a una conclusión, se procederá de la siguiente manera: En primer lugar, se realizarán las pruebas necesarias para determinar la estacionariedad de los conjuntos de datos. Una vez obtenidos los conjuntos con estas características, daremos lugar al análisis de la normalidad utilizando, posteriormente, el porcentaje de falsos descubrimientos (FDR) para obtener una mayor exactitud del resultado.

Abstract

The aim of this study is to find whether the heights of the sea waves follow a Gaussian process. The study of the observations stored in 11 different buoys around the world will be carried out. To do this, 2305 different consecutive observations have been selected in each buoy between 14h and 15h on 23 July 2018. It is convenient to know the distribution of this phenomenon because many of the elements that are in the sea such as oil infrastructure, breakwaters or even the ships themselves, require a pre-construction modeling to determine, for example, each resistance. In order to reach a conclusion, we will proceed as follows: Firstly, the necessary tests will be carried out to determine the stationarity of the data sets. Once the sets with these characteristics have been obtained, we will continue with the normality analysis using, subsequently, the false discovery rate (FDR) to obtain a greater accuracy of the result.

Keywords: Gaussianidad, Proyección aleatoria, False Discovery Rate.

1 Introducción

En matemáticas, los test de bondad de ajuste son utilizados comunmente para determinar si un conjunto de datos son realizaciones de una distribución normal. En este caso, se va a estudiar la distribución que siguen las alturas de las olas del mar utilizando las mediciones obtenidas de las boyas ubicadas en determinadas estaciones marítimas. Un proceso estocástico es un conjunto de variables aleatorias sobre las que se realizan observaciones y, los valores observados, forman la serie temporal. Por tanto, se puede decir que cada una de las boyas que se va a estudiar es un proceso estocástico.

Hasta ahora, de las investigaciones y estudios realizados de la distribución de las olas del mar cuando este está en calma, se ha obtenido que el fenómeno natural formado sigue una distribución Gaussiana. Es importante hacer hincapié en que, para el estudio, los datos deben ser obtenidos cuando el mar no está influido por agentes externos. De ser así, nuestra serie no sería estacionaria y, por tanto, no podríamos utilizar los test que estudian la Gaussianidad de una serie temporal ya que están basados en series con efecto estacional. El estudio más común es el del conjunto finito de datos modelado como procesos estacionarios de segundo orden, esto significa que la media y la varianza se mantienen constantes en el tiempo. Además, es común asumir la hipótesis de Gaussianidad, pues esta característica asigna al proceso propiedades muy beneficiosas a la hora de realizar cálculos estadísticos o predicciones. Cuando un proceso sigue una distribución normal, entonces el proceso que habíamos catalogado como proceso estacionario se convierte en un proceso estrictamente estacionario.

El estudio de variables aleatorias en $\mathbb{R}^p$ no es demasiado complejo. Por ejemplo, supongamos que estamos en $\mathbb{R}$. Imaginemos que se tiene una variable aleatoria X y que obtenemos una muestra de resultados de esa variable aleatoria $x_1, \ldots, x_n$. Si es normal, esta variable queda únicamente determinada por su media μ y su desviación típica σ y, al ser unidimensional, se estudiaría si la variable en $\mathbb{R}$ sigue una normal $N(\mu, \sigma^2)$. Análogamente ocurre cuando estamos en $\mathbb{R}^2$. Imaginemos que se tiene de nuevo una variable aleatoria X pero esta vez en $\mathbb{R}^2$ y que se obtiene una muestra de resultados $x_1, \ldots, x_n$ de la variable aleatoria. En este caso, se comparan las propiedades de nuestro conjunto de resultados con las propiedades de una distribución $N\left[\begin{pmatrix}\mu_x & \mu_y\end{pmatrix}^T, \begin{pmatrix}\sigma_x^2 & \sigma_{xy}^2; \sigma_{yx}^2 & \sigma_y^2\end{pmatrix}\right]$, donde σ_{xy} es la covarianza entre las variables X e Y. De igual forma, sucedería con dimensiones más grandes. Para poder saber la distribución que siguen las variables aleatorias en $\mathbb{R}^p$ son comunes los test como el de Kolmogorov-Smirnov, que calcula la distancia vertical máxima entre las funciones de distribución acumulada empírica de dos muestras o entre una función de distribución acumulada de una muestra y la de la teoría de referencia. Otro tipo de tests que estudian la Gaussianidad de un conjunto de datos en $\mathbb{R}^p$ son Shapiro-Wilks, Lilliefors y Anderson-Darling entre otros. Sin embargo, en este proyecto se busca estudiar la distribución del proceso estocástico estacionario X y, para ello, se obtendrá una muestra de este proceso que seguirá siendo de dimensión infinita. Para su estudio, son conocidos los test de Epps y Lobato y Velasco que miran si la variable aleatoria unidimensional, X_i, es normal. No obstante, para saber si un proceso es normal, lo que hay que mirar es que cada vector finito-dimensional sea normal. Por tanto, para saber realmente la distribución del proceso estocástico, no se estudia únicamente la distribución de las X_i, si no la distribución de cada vector $(X_1, \ldots, X_k)$ para todo k.

Por ello, en este trabajo se va a utilizar una prueba para el estudio de la Gaussianidad de este tipo de conjunto de datos. Esta prueba va a tener como base el estudio de la Gaussianidad centrado en el test de hipótesis del proceso estacionario de variables aleatorias de valores reales $X := (X_t)_{t\in\mathbb{Z}}$ con las siguientes hipótesis:

$$H_0 : \text{X es Gaussiano} \qquad H_a : \text{X no es Gaussiano} \tag{1}$$

Como hemos comentado antes, existen test importantes basados en las características de la distribución Gaussiana como lo son, por ejemplo:

- El test de Epps (1987), basado en la función característica.
- El test de Lobato y Velasco (2004), basado en la asimetría (o sesgo) y la curtosis.

- El test de Moulines and Choukri (1996) que estudia conjuntamente la función característica, la asimetría y la curtosis del proceso.

Estos tests estudian la Gaussianidad de marginales unidimensionales de un proceso. Sin embargo, considerar que un proceso es Gaussiano estudiando únicamente su marginal unidimensional, nos puede dar lugar a error, ya que existen procesos no Gaussianos con marginales unidimensionales Gaussianas. Aquí, lo que se quiere expresar es que a la hora de probar la hipótesis de Gaussianidad con Epps y con Lobato y Velasco obtendríamos un error de tipo I del cual se hablará en la Sección 4. A pesar de ello, los tests servirán como introducción al estudio que se va a realizar sobre nuestro proceso estocástico (determinar si la altura de las olas del mar sigue una distribución normal mediante pruebas de hipótesis múltiples con proyección Gaussiana).

Como ya se ha comentado anteriormente, interesa realizar el estudio de una serie temporal infinita, por ello, se han recogido las mediciones de la altura de las olas de mar mediante boyas que van a permanecer en la misma ubicación de manera indefinida.

Antes de hablar en detalle de los test de gaussianidad mencionados, se van a dar una serie de conceptos que serán necesarios para un mejor entendimiento de los procedimientos realizados a lo largo del proyecto. Empezaremos profundizando sobre el concepto de *serie temporal o proceso estocástico*. La necesidad de su comprensión es importante debido a que, como hemos introducido anteriormente, el estudio se basa en el análisis de este tipo de conjuntos de datos. Por otra parte, se realizará un recordatorio del denominado *test de hipótesis* y se enunciarán los distintos métodos utilizados para el análisis de los datos finalizando con la Sección 5 en el que se incluyen los resultados obtenidos al realizar el estudio correspondiente a las mediciones de las olas del mar.

2 Preliminares

Antes de introducir el concepto de serie temporal, recordemos algunas nociones.
Sean dos espacios medibles (Ω, σ) y (Ω^*, σ^*). Sea Ω un conjunto no vacío y $\sigma \subset P(\Omega)$, σ es una σ-álgebra si cumple las siguientes condiciones:

1. $\Omega \in \sigma$
2. Si $A \in \sigma$, entonces $A^c \in \sigma$, donde A^c representa el conjunto complementario de A.
3. Si $\{A_n\}_{n=1}^{\infty} \in \sigma$, entonces $\cup_{n=1}^{\infty} A_n \in \sigma$

Entonces, una **variable aleatoria** [8] es una aplicación $X : \Omega \longrightarrow \Omega^*$ tal que, $X^{-1}(B) \in \sigma$ para todo $B \in \sigma^*$. Además, si cumple que $\sigma^* = \mathbb{R}$ y $\sigma^* = \beta$ donde β es la σ-álgebra de Borel se dice que X es una **variable aleatoria real**. De manera sencilla, una variable aleatoria es una función cuyos valores son los resultados de un experimento aleatorio.

Para un mayor entendimiento del proyecto, definamos adicionalmente los siguientes conceptos relacionados con la noción de proceso estocástico. Sea $\{X_t : t \in \mathbb{Z}\}$ un proceso estocástico[14]:

Definición 2.1 *Se define la* ***función de distribución de primer orden*** *de* $(X_t)_{t\in\mathbb{Z}}$ *como:* $F_X(x,t) = P(X_t \leq x)$ *y, por tanto, para una distribución absolutamente continua, se tiene también la función de densidad de primer orden derivando la función de distribución respecto a x.*

$$f(x,t) = \frac{dF_X(x,t)}{dx}$$

Definimos de la misma manera el siguiente concepto:

Definición 2.2 *Se define como* ***función de distribución de segundo orden*** *del proceso* $(X_t)_{t\in\mathbb{Z}}$ *como:* $F_X(x_1, x_2, t_1, t_2) = P(X_{t_1} \leq x_1 \cap X_{t_2} \leq x_2)$ *y, adicionalmente, se tiene la función de densidad de segundo orden derivando parcialmente respecto a* x_1 *y a* x_2 *la función de distribución .*

$$f(x_1, x_2, t_1, t_2) = \frac{\partial^2 F(x_1, x_2, t_1, t_2)}{\partial x_1 \partial x_2}$$

Definición 2.3 ***Distribución conjunta:*** *La distribución conjunta muestra la distribución de probabilidad de dos o más variables. La definición formal para variables aleatorias discretas es la siguiente:* $D_{XY}(x, y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}((X = x) \cap (Y = y))$. *La finalidad es buscar las relaciones entre la dos variables.*

Recordemos que la suma de las probabilidades de las variables es 1, es decir:

$$\sum_{(x,y) \in R_{XY}} \mathbb{P}(X = x, Y = y) = 1$$

Definición 2.4 ***Distribución Marginal:*** *Llamamos distribución marginal de X de la distribución conjunta de XY, a la distribución de X obtenida de la función de distribución conjunta de X e Y. Esto es, fijado un valor de X, obtenemos las probabilidades del valor de X con respecto a cada valor de Y y viceversa. Lo escribimos de la siguiente manera:*

$$\mathbb{P}_X(x) = \mathbb{P}(X = x) = \sum_{y \in R_Y} \mathbb{P}(X = x, Y = y_j) = \sum_{y \in R_Y} \mathbb{P}_{XY}(x, y_j), \text{ para cualquier } x \in R_x$$
$$\mathbb{P}_Y(y) = \mathbb{P}(Y = y) = \sum_{x \in R_X} \mathbb{P}(X = x_j, Y = y) = \sum_{x \in R_X} \mathbb{P}_{XY}(x_j, y) \text{ para cualquier } y \in R_y$$

Definición 2.5 ***Distribución finito-dimensional****: La distribución finito dimensional de* $(X_t)_{t \in \mathbb{Z}}$ *son las funciones de distribución conjuntas de* $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$, $t_1, t_2, \ldots, t_n \in T$, $n \in \mathbb{N}$

Además, analizando el comportamiento del proceso estocástico obtenemos que[17]:

- El proceso es estable en media (o de primer orden) si $\mu_t = \mu = cte$.
- El proceso es estable en varianza si $\sigma_t^2 = \sigma_x^2 = cte$.
- El proceso es estable en autocovarianza si $Cov(t, s) = Cov(s, t) = Cov(X_t, X_s)$.
- El proceso es estacionario débil (o de segundo orden) si tanto la media como la varianza son constantes en el tiempo.
- El proceso es estacionario en el sentido estricto si las distribuciones marginales de todas las variables son idénticas y, además, la distribución finito dimensional de cualquier conjunto de variables sólo depende de los retardos. Es decir, si:

  $$F_{t_1, \ldots, t_k}(x_1, \ldots, x_k) = F_{t_{1+h}, \ldots, t_{k+h}}(x_1, \ldots, x_k)$$

  para cualquier $k \in \mathbb{N}, t_1, \ldots, t_k, h \in \mathbb{R}$, donde $F_{t_1, \ldots, t_k}$ denota la distribución conjunta de $X_{t_1}, \ldots, X_{t_k}$.
- El proceso se dice que es de segundo orden si $\mathbb{E}[|X_t|^2] < \infty$ para todo $t \in T$.

Definición 2.6 *La* ***estacionariedad de orden N*** *es un tipo de estacionariedad débil donde se requiere que la distribución de las n muestras del proceso estocástico debe ser igual a la distribución de las muestras desplazadas en el tiempo para todas las muestras n hasta un cierto orden N. Un proceso aleatorio* $(X_t)_{t \in \mathbb{Z}}$ *se dice que es estacionario de orden N si:*
$F_X(x_{t_1+\tau}, \ldots, x_{t_n+\tau}) = F_X(x_{t_1}, \ldots, x_{t_n})$ *para todo* $\tau, t_1, \ldots, t_n \in \mathbb{R}$ *y para todo* $n \in \{1, \ldots, N\}$

Cabe destacar que una propiedad muy útil es la condición de Gaussianidad en un proceso estocástico estable, ya que es el único caso en el que proceso estacionalmente débil implica, a su vez, proceso estacionario en el sentido estricto.

2.1 Series temporales y procesos estocásticos

Una **serie temporal** es un conjunto de observaciones de una variable obtenidas secuencialmente en el tiempo de manera equiespaciada [22]. Lo denotaremos por X_{ik} donde, por ejemplo, i se refiere al año y k al momento del año en el que se obtiene la observación. Las series temporales también pueden considerarse como un caso particular de las variables estadísticas bidimensionales *(A, B)* donde la variable independiente, A, es el tiempo y la variable dependiente, B, es la variable cuya distribución temporal se pretende analizar. Introduzcamos ahora la noción de proceso estocástico. Un **proceso** es un conjunto de observaciones obtenidas secuencialmente en el tiempo. Si este conjunto de datos son variables aleatorias obtenidas de manera equiespaciada tal que $\{X_t : t \in \mathbb{Z}\}$, entonces lo denominamos **proceso estocástico** [17]. En términos mucho más sencillos, un proceso estocástico es aquel que no se puede predecir. Se mueve al azar. Aunque, como veremos más tarde, existen distintos tipos de procesos estocásticos.

La manera en la que se debe de pensar sobre los procesos estocásticos es comparándolo con procesos deterministas. Un proceso determinista es aquel del que se puede hallar exactamente su valor futuro sin necesidad de involucrar a la aletoriedad. Un ejemplo claro y sencillo de proceso determinista es el paso de una medida a otra: Si se calcula el número de litros que son $300cm^3$ el resultado será $0,3l$, y no habrá posibilidad de que sea otro valor. Son fórmulas exactas que, dado un valor, se obtendrá otro valor determinado y será siempre el mismo. El proceso estocástico difiere de lo anterior en que, en cada paso se va a tener aleatoriedad, no se sabe dónde se va a estar, pero lo que sí que sabes es que hay alguna distribución de X_t en ese momento. Puntualizar que cada una de las variables aleatorias que componen el proceso estocástico tendrá una distribución.

La relación entre la serie temporal y el proceso estocástico que la genera, es análoga a la que existe entre una muestra y la población de la que procede, de tal forma que podemos considerar una serie temporal como una muestra o realización de un proceso estocástico formado por una sola observación de cada una de las variables que componen el proceso.

Definamos entonces de manera formal las definiciones de estos dos conceptos en las que se considera la relación existente entre cada uno[17]:

Definición 2.7 (Proceso estocástico) *Un proceso estocástico es un conjunto de v.a's (X_t) tal que $t \in C$. Llamamos trayectoria del proceso a una realización del proceso estocástico. Si C es discreto, el proceso es en tiempo discreto, mientras que si C es continuo, el proceso es en tiempo continuo.*

Definición 2.8 (Serie temporal) *Una serie temporal es una realización de un proceso estocástico en tiempo discreto donde los elementos de C están ordenados y corresponden a instantes equidistantes del tiempo.*

Ejemplo 1. Sea una boya, A, de la cuál se van a obtener 3 muestras distintas de las mediciones de la altura de las olas del mar. Se considera que esta boya recoge datos de manera continua. Se va a seleccionar muestras formadas por 50 observaciones obtenidas de la boya A a las 17h durante tres días consecutivos bajo las mismas condiciones.

Serie Temporal: Cuando se tome el primer día la muestra correspondiente, es decir, cuando se obtenga el primer conjunto de observaciones de las medidas de la altura de las olas, se obtendrá una función de una sola variable. Los valores que obtiene dicha función, forman la denominada serie temporal. Análogamente sucederá para las muestras obtenidas de los días 2 y 3. Por tanto, en la *Figura 1*, se tiene tres series temporales.

Proceso estocástico: Para poder crear un proceso estocástico denotemos en primer lugar a su realización como $X(\omega)$. En nuestro caso tendremos tres secciones de realizaciones, una por muestra obtenida al observar a lo largo del tiempo nuestro conjunto de variables aleatorias tal que $X_t(\omega)$ con t=1,2 y 3. Entonces denotaremos al proceso estocástico como $X(X_t(\omega), t)$ que, segun la magnitud que se fije, se obtendrá o una variable aleatoria, o una función de una sola variable. Si fijamos el momento del tiempo t, entonces estaremos creando una variable aleatoria cuyos valores serán el de la altura de las olas en cada una de las realizaciones. Por otra parte, si fijamos la realización $X_t(\omega)$ y dejamos

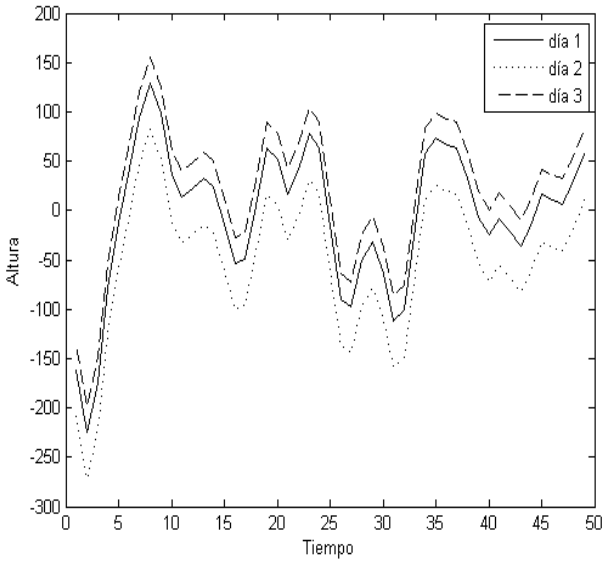

Figura 1: Series temporales obtenidas de la boya A durante 3 días consecutivos

como variable el tiempo, obtendremos una función de una sola variable en función del tiempo que se corresponderá con la realización que hayamos fijado. Entonces obtenemos en la *Figura 2*, las tres secciones de muestras de observaciones obtenidas de nuestro proceso estocástico en la que se identifica claramente que el valor de la altura de las olas del mar dependen del tiempo y de la realización.

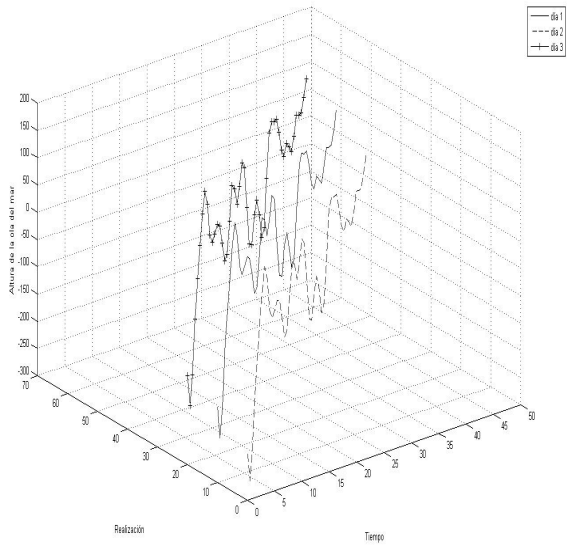

Figura 2: Proceso estocástico formado a partir de mis datos

En resumidas cuentas, el proceso estocástico se puede considerar como un conjunto de variables aleatorias si se fija la variable del tiempo t, o un conjunto de funciones si se fija la realización $X_t(\omega)$.

2.2 Características de las series temporales

Hablemos ahora de las características y propiedades de los procesos estocásticos. Dado que una serie temporal es una muestra o realización del proceso estocástico, podemos extrapolar las características de la serie temporal a dicho proceso.

Una de las característica principales de este tipo de series es su no independencia, es por eso por lo que es posible realizar la predicción del siguiente valor que le corresponde a la serie. Teniendo en cuenta esto, las series temporales se pueden clasificar, según su predicción, en dos tipos [17]: *Deterministas y Estocásticas.*

- **Definición 2.9 (Serie determinista)** *Una serie se dice determinista si se pueden predecir exactamente los valores a partir de las observaciones obtenidas.*
- **Definición 2.10 (Serie estocástica)** *Se llaman series estocásticas aquellas series que únicamente se pueden predecir de manera parcial a raíz de las observaciones pasadas y no se pueden determinar exactamente. Se considera que los futuros valores tienen una distribución de probabilidad que está condicionada por los valores pasados.*

Cuando analizamos una serie temporal, es importante saber cuál es el comportamiento de dicha serie, para ello, podemos determinar cada uno de sus componentes. Son componentes de una serie temporal la tendencia, la estacionariedad y el ruido blanco. Introduzcamos brevemente cada uno de los conceptos.

2.2.1 Tendencia

Cuando hablamos de la ***tendencia*** de una serie, hablamos de cómo van evolucionando los datos a lo largo del tiempo. Es decir, se identifica con el cambio a largo plazo de la media. Podremos encontrarnos con una tendencia ascendente o descendente. Para poder determinar la tendencia, lo más común es suponer que la serie no es estacionaria *(ver Sección 2.2.2)* y, por lo tanto, $E_t = 0$. Así, posteriormente, se podrá realizar test de hipótesis sobre la serie temporal. La tendencia puede ser *determinística o evolutiva.*

- **Definición 2.11** *Se llama **tendencia determinística** a la tendencia que puede ser determinada con modelos de regresión lineales simples, polinomios, curvas, etc.*

  Por ejemplo, podemos calcular la tendencia mediante el modelo de regresión lineal: $T_t = a + bt$ donde a y b son los coeficientes de la recta. De esta manera, tendríamos $E_t = 0$ y $T_t = a + bt$, y entonces $I_t = X_t - \hat{a} - \hat{b}_t$ *(ver secciones 3.2.2 y 3.2.3 respectivamente).*

- **Definición 2.12** *La **tendencia evolutiva** es un tipo de tendencia que utiliza la media móvil para poder ser determinada.*

  Como se ha comentado anteriormente, se puede observar una tendencia creciente o tendencia decreciente: Sean X_{t-1}, X_t y X_{t+1} datos consecutivos de nuestra serie temporal, entonces tendremos tendencia evolutiva si se da, por ejemplo:

  - $X_{t-1} = X_t - \delta$
  - $X_{t+1} = X_t + \delta$

  donde δ es el paso o crecimiento.

A continuación, un ejemplo de serie que presenta tendencia[1]:

2.2.2 Estacionariedad

Una serie tiene **efecto estacional** cuando es estable a lo largo del tiempo sin que se aprecien aumentos o disminuciones sistemáticos de sus valores. En términos matemáticos, esto sucede cuando la media y la varianza son constantes a lo largo del tiempo. Gracias a la propiedad de que la media es constante ($E[X_t] = cte$), si se desea realizar el estudio, se puede estimar la media y utilizar el dato obtenido para el cálculo de la predicción del dato siguiente en la serie. Por otra parte, cuando una serie es estacionaria, se pueden obtener intervalos de confianza para las predicciones, asumiendo que X_t sigue una distribución conocida.

[1]Los datos han sido obtenidos de 'http://verso.mat.uam.es/ joser.berrendero/datos/gas6677.dat'

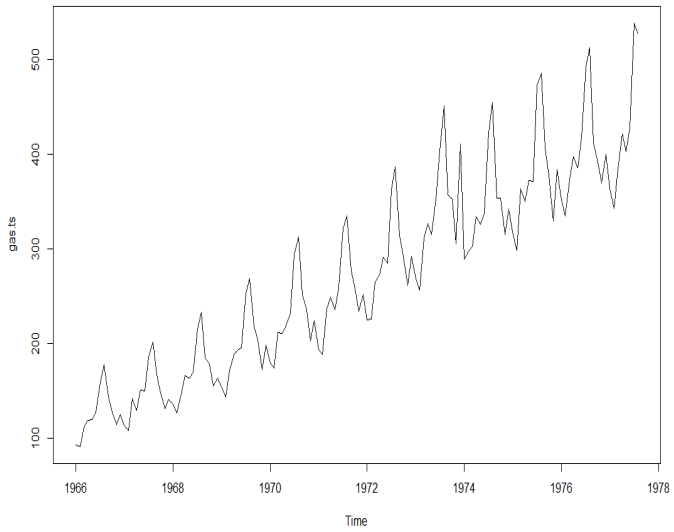

Figura 3: Serie con tendencia ascendente

Los modelos de predicción de series temporales están creados para ser utilizados en series estacionarias [2], por ello merece la pena mencionar que siempre es posible transformar una serie no estacionaria en una estacionaria para poder predecir el siguiente dato de la serie aplicando los mismos métodos que se podrían aplicar en una serie inicialmente estacionaria. Si se va a realizar este estudio, no hay que olvidarse de realizar de nuevo el camino inverso para obtener los datos que queríamos de la serie original.

Con carácter general, queremos que una serie temporal tenga componente estacional, es decir, que no tenga tendencia. La estacionariedad normalmente es una propiedad de un proceso estocástico, no de una serie temporal, pero decimos serie temporal estacionaria si se piensa que puede ser modelizada con modelos estacionarios o procesos estocásticos estacionarios.

Por tanto, en una serie temporal estacionaria no hay:

- Cambio sistemático en la media
- Cambio sistemático en la varianza
- Variaciones periódicas con periodo superior al año

En la *Figura 4*, se puede observar un ejemplo de serie estacionaria.

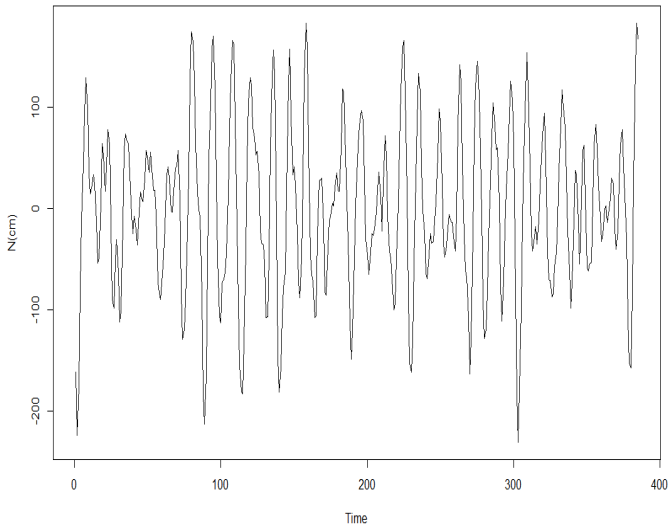

Figura 4: Serie con componente estacional

2.2.3 Componente aleatoria o ruido blanco

En tercer lugar, es muy común encontrarse con ciertas observaciones que no siguen ningún criterio después de haber identificado y tras haber eliminado las componentes calculadas anteriormente de la serie (tendencia y estacionariedad). Para analizar estas observaciones, trataremos de estudiar qué tipo de comportamiento aleatorio presentan estos residuos mediante algún modelo probabilístico que los describa. Denominaremos a esta componente de la serie *componente aleatoria o ruido blanco*.

El **ruido blanco** [2] es el proceso estocástico en el que las variables aleatorias que lo forman no están conectadas entre sí siendo $E[X_t] = 0$ y $Var(X_t) = \sigma^2 = cte$.

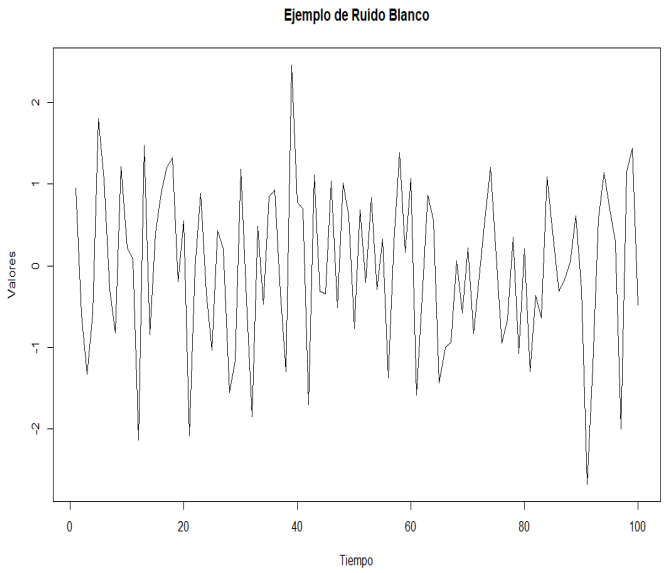

Figura 5: Ruido blanco o Componente aleatoria

Teniendo en cuenta toda la información obtenida hasta ahora, será fácil comprender las siguientes características de una serie temporal:

- El orden de los datos importa

- Las observaciones de una serie temporal no son independientes

Esto nos lleva a introducir el concepto de *dependencia temporal*. Se dice que una serie tiene dependencia temporal cuando los datos del pasado afectan al valor de la variable en el futuro. Para analizar series temporales es común utilizar modelos de regresión y, a mayor cantidad de datos se cojan para la observación, mayor calidad del resultado obtendremos. Por ello, a la hora de realizar un análisis de una serie temporal es importante tener en cuenta los siguientes factores:

- Saber datos y entender el pasado
- Entender cuál es la situación actual
- Predecir el dato futuro teniendo en cuenta el error de predicción

Teniendo en cuenta estos factores, matemáticamente podríamos descomponer la serie de la siguiente manera:

$$X_t = T_t + E_t + I_t \tag{2}$$

Donde T es la tendencia en el momento t, E la parte estacional e I la componente aleatoria.

3 Metodología

3.1 Proyección aleatoria

Para dar lugar a nuestro análisis del proceso estocástico vamos a introducir otro concepto, la denominada: *Proyección aleatoria (RP)*. La proyección aleatoria es una técnica utilizada en el campo de las matemáticas para reducir la dimensión de un conjunto de datos que se encuentra en el espacio euclideo. Para ello se utiliza una matriz aleatoria cuyas columnas tienen vectores de norma uno. Aunque este método ha atraído mucho interés, los resultados empíricos son escasos[5].

Dado $\{X_t\}_{t\in\mathbb{Z}}$ el proceso estocástico, se va a tomar un vector aleatorio de dicho conjunto que seguirá siendo infinito tal que $(X_p)_{p<t}$. El resultado de realizar un test de hipótesis cuando se aplica la proyección aleatoria a un vector es un número. Esto quiere decir que, al reducir de dimensión, se va a perder determinada información, pero se seguirá teniendo mucha del proceso estocástico $\{X_t\}_{t\in\mathbb{Z}}$. Sin embargo, si realizamos varias veces el test de hipótesis y, por ende, aplicamos varias veces la proyección aleatoria, en cada iteración perdemos información distinta. Por ello, si se analizaran las proyecciones en conjunto, se obtendría un conjunto de datos rico en información del proceso inicial.

El teorema que vamos a enunciar a continuación viene de los papeles de [19] y [9] en los que se estudia las proyecciones aleatorias basadas en los test de Gaussianidad, por lo que podríamos utilizarlo para saber si el conjunto de datos que estamos estudiando es o no Gaussiano. En él, se hace referencia al término distribución disipativa. Introduzcamos este concepto:

Definición 3.1 (Distribución disipativa) *Sea D un elemento aleatorio del espacio de Hilbert $\mathbb{H}$. Diremos que su distribución es disipativa si cumple lo siguiente:*
• *$\exists$ una base ortonormal $\{v_n\}_{n=1}^{\infty}$ de $\mathbb{H}$, tal que $\mathbb{P}(D_{V_n^{\perp}} = 0) = 0$, para todo $n \geqslant 2$*
• *La distribución condicionada de D_{V_n} dado $D_{V_n^{\perp}}$ es absolutamente continua con respecto la medida de Lebesgue n-dimensional.*

Teorema 3.1 *(Cuesta-Albertos et al., 2007 [9]).*
Si η es una distribución disipativa en $\mathbb{H}$ y $D = (X_1, ..., X_t)$ un elemento aleatorio de $\mathbb{H}$, entonces es Gaussiano si y solo si $\eta(E) > 0$ donde E= $\{h \in \mathbb{H}$: la distribución de $\langle D, h\rangle$ es Gaussiana$\}$.

Este resultado es relevante debido a que, si η es una función disipativa, se cumplen las siguientes propiedades:
• $\eta(E) \in \{0, 1\}$

- D no es Gaussiano si y solo si $\eta(E) = 0$
- D es Gaussiano si y solo si $\eta(E) = 1$

En resumen, si queremos saber si una distribución de D es Gaussiana, entonces vamos a seleccionar un punto aleatorio h $\in \mathbb{H}$ utilizando la función disipativa y ver si el valor real de la variable aleatoria $\langle D, h\rangle$ es Gaussiana. De este resultado obtenemos que:

- Un test para calcular la Gaussianidad a nivel α de una proyección unidimensional elegida aleatoriamente es, además, un test al mismo nivel para probar la Gaussianidad del proceso X.
- Un test de Gaussianidad consistente aplicado a la proyección, es un test consistente para la Gaussianidad de todo el proceso X.

Este último punto nos resulta de gran importancia, pues es una propiedad utilizada en *Cuesta-Albertos et al.,(2007) [9]* para construir un test Gaussiano dada una muestra aleatoria de trayectorias. Es cierto que, tras realizar la proyección, nosotros tendremos una secuencia de observaciones extraídas de una trayectoria fija pero, dado que el teorema anterior transforma el análisis de la Gaussianidad del proceso X en el análisis de la Gaussianidad de una proyección unidimensional elegida aleatoriamente, aplicando el teorema, podremos utilizar la teoría de *Cuesta-Albertos et al.,(2007)[9].*

3.2 Test de hipótesis

Cuando se realizan estudios y se interpretan hallazgos, los investigadores deben evaluar si los resultados han ocurrido por casualidad. Esto da lugar a los test de hipótesis. El test de hipótesis es un procedimiento sistemático que se utiliza para decidir si el resultado de un estudio apoya a una determinada teoría que ha sido aplicada a una población. Para ello, los datos utilizados son muestras extraídas de la población total y se extrapola el resultado.

El contraste de hipótesis está formado por la hipótesis nula y la hipótesis alternativa [18]. La *hipótesis nula* H_0 es la hipótesis que asume que no hay diferencia, asociación o relación entre las variables. La *hipótesis alternativa* H_1 (denominadas por H_a) es la hipótesis que sugiere que las observaciones de la muestra están influenciadas por una causa no aleatoria. Asume diferencia, asociación o relación entre las variables. La redacción específica de la hipótesis alternativa es importante ya nos dice si necesitamos realizar un test de una o dos colas (one-tailed o two-tailed).

El **test de una cola** resulta de una hipótesis alternativa que especifica una dirección, es decir, cuando la hipótesis alternativa afirma que el parámetro es de hecho mayor (de cola derecha) o menor (de cola izquierda) que el valor especificado en la hipótesis nula.

El **test de dos colas** resulta de una hipótesis alternativa que no especifica una dirección, es decir, cuando la hipótesis alternativa afirma que la hipótesis nula es errónea.

La principal diferencia entre las pruebas de una cola y las de dos colas es que las pruebas de una cola sólo tendrán una región crítica[2] mientras que las de dos colas tendrán dos regiones críticas. Si requerimos un intervalo de confianza[3] del $100(1-\alpha)\,\%$ tenemos que hacer algunos ajustes al usar una prueba de dos colas. El intervalo de confianza debe permanecer de tamaño constante, así que si realizamos una prueba de dos colas, las regiones críticas deben tener la mitad del tamaño, pues tendremos dos regiones en lugar de una sola. Esto significa que cuando leemos las tablas, al realizar una prueba de dos colas, tenemos que considerar $\frac{\alpha}{2}$ en lugar de α.

Para la realización de un test de hipótesis se utiliza el estadístico del test. Esto es un valor que resume todo el conjunto de datos y la elección del mismo variará dependiendo de la distribución que se utilice. Si el estadístico se encuentra en la región crítica, entonces se acepta la hipótesis alternativa. En caso contrario, se acepta la hipótesis nula. El p-valor es el resultado obtenido del estadístico. Entonces, una decisión entre dos hipótesis se realiza comparando el p-valor con el valor de significación[4], que es la probabilidad u oportunidad de tener los datos o población bajo las condiciones de la hipótesis nula.

[2]Una región crítica o región de rechazo es un conjunto de valores para el estadístico del test para el cual se rechaza la hipótesis nula

[3]Un intervalo de confianza o región de aceptación, es un conjunto de valores del estadístico del test para el que se acepta la hipótesis nula

[4]Punto de corte para determinar si rechazar o aceptar la hipótesis nula

Suponiendo que la hipótesis nula es cierta, si el p-valor es inferior al nivel de significación, α, entonces se acepta la hipótesis alternativa. Si rechazamos la hipótesis nula en un nivel de significación α_1 pero aceptamos la hipótesis nula en un nivel de significación α_2 con $\alpha_1 > \alpha_2$ entonces sabemos que el p-valor está entre α_1 y α_2.

Entonces, ante un test de hipótesis concluiremos que:

• Si el p-valor es menor que el nivel alfa ($p - valor < \alpha$) , rechazaremos la hipótesis nula y, cuanto más pequeño sea, más fuerte será la evidencia de que la hipótesis nula debe ser rechazada.

• Si el p-valor es mayor que el nivel alfa ($p - valor > \alpha$), no tendremos evidencias suficientes para rechazar la hipótesis nula, lo que hace que la prueba sea no concluyente.

3.3 Métodos de estacionariedad e independencia

El estudio de la independencia y de la estacionariedad es importante ya que la teoría existente es para procesos con componente estacional. Como ya enunciamos en la Sección 2, un proceso estacionario es un proceso estocástico cuya distribución de probabilidad en un instante de tiempo fijo o una posición fija es la misma para todos los instantes de tiempo o posiciones. Para determinar si nuestros conjuntos son estacionarios se van a realizar los test siguientes:

- **Box-Pierce y Ljung-Box test:**
  Los modelos de las series temporales, en concreto, ARMA (autoregresive moving average), pueden ser vistos como transformadores de los datos en ruido blanco. Si el modelo se ha elegido correctamente, habrá cero autocorrelación en los errores, es decir, no habrá dependencia entre los errores. En 1970, Box y Pierce[6] propusieron un estadístico (Q_{BP}) para determinar si la autocorrelación en una serie temporal débilmente estacionara es distinta de cero.

  Recordatorio 3.1 *Recordemos de manera fugaz que la correlación, ρ_{xy}, es el valor que determina la existencia de dependecia o no entre dos variables x e y. Si se obtiene que $\rho_{xy} > 0$ se tendrá que existe una dependencia directa entre las dos variables y, por el contrario, con $\rho_{xy} < 0$ se concluirá que existe una dependencia inversa o negativa. Finalmente, no existirá relación lineal entre dos variables si $\rho_{xy} = 0$.*

  Para probar la independencia de esta serie, se apoyó en el test con las siguientes hipótesis:

  H_0: Muestra con autocorrelación cero
  H_a: Muestra con autocorrelación distinta de cero

  cuyo estadístico es

  $$Q_{BP} = n \sum_{j=1}^{m} \rho(j)^2$$

  donde n es el número de observaciones de la serie temporal, m el número de correlaciones que se quiere obtener y donde ρ el coeficiente de correlación:

  $$\rho(j) = \frac{\hat{\gamma}(j)}{\hat{\gamma}(0)}$$

  con $\hat{\gamma}(0)$ el estimador de la autocovarianza tal que:

  $$\hat{\gamma}(j) = (n-j)^{-1} \sum_{t=1}^{n-j} [(y_t - \hat{\mu})(y_{t+j} - \hat{\mu})]$$

  concluyendo que, si eran independientes e identicamente distribuidas, entonces se cumplía que $Q_{BP} \sim \chi^2_m$ donde m son los grados de libertad.

  No obstante, para aumentar la fuerza a este test para muestras finitas y cuando se estudia la independencia de series temporales no normales, en 1978 Ljung-Box modificó este estadístico obteniendo así mejores resultados:

$$Q_{LB} = n(n+2)\sum_{j=1}^{m}\frac{\rho(j)^2}{n-j}$$

El estadístico Q_{LB} se distribuye asintóticamente como χ^2 con m grados de libertad. Entonces, rechazaremos la hipótesis nula de independencia cuando $\chi^2_m > Q_{LB}$.

- **Augmented Dickey-Fuller test:**
  La prueba ADF[10] evalua si existen raices unitarias para determinar la estacionariedad de una serie temporal. Por raiz unitaria se entiende la característica de una serie temporal que la hace no estacionaria. Técnicamente hablando, una raiz unitaria existe en una serie temporal si el valor de α en la ecuación siguiente es igual a 1:

$$Y_t = \alpha Y_{t-1} + \beta X_e + \epsilon$$

donde Y_t es el valor de la serie temporal en el momento t, X_e la variable exógena, esto es, la variable cuyo valor está determinado por factores externos al modelo en el que se incluye y ϵ es el ruido blanco de la serie temporal. Entonces hablemos del test ADF que tiene la siguiente hipótesis:

H_0: Muestra no estacionaria
H_a: Muestra estacionaria

El ADF test es una extensión del test Dickey Fuller (DF), por ello, se va a introducir este test. Como se ha dicho anteriormente, es un test basado en observar la raiz unitaria de los modelos de regresión. En este caso, DF presentó el siguiente modelo como hipótesis nula:

$$Y_t = \mu + \beta t + \alpha y_{t-1} + \Phi Y_{t-1} + e_t$$

Como se puede observar, tiene una hipótesis nula similar al del test de la raiz unitaria, esto es, si el coeficiente de Y_{t-1} es 1 esto implica la presencia de una raiz unitaria. Si no se rechaza, la serie entonces se considera no estacionaria. Ahora bien, el ADF test, involucra a la ecuacion anterior y es uno de los más frecuentes para el cálculo de la raiz unitaria. Obviamente, el ADF test se basa en el DF test y amplía la ecuación para incluir procesos regresivos de ordenes más altos, tenemos entonces la siguiente ecuación:

$$Y_t = \mu + \beta t + \alpha y_{t-1} + \sum_{j=1}^{p}\Phi_j \Delta Y_{t-j} + e_t$$

Si nos fijamos, solo hemos añadido más términos de diferencias, mientras que el resto de la ecuación sigue igual y, la hipótesis nula, sigue siendo la misma que en el DF test.

El estadístico DF es calculado como:

$$ADF = \hat{\sigma}/SE(\hat{\sigma})$$

Donde $\hat{\sigma}$ es el coeficiente de estimación y $SE(\hat{\sigma})$ la estimación correspondiente del error estandar para cada tipo de modelo lineal. Un punto clave para recordar es que, dado que la hipótesis nula asume la presencia de raices unitarias, esto es $\alpha = 1$, el p-valor obtenido para poder concluir que la serie es estacionaria debe ser menor que el nivel de significación (0.05) y así poder rechazar la hipótesis nula de no estacionariedad.

- **Kwiatkowski-Phillips-Schmidt-Shin test:**

  KPSS [15] determina si una serie temporal es estacionaria alrededor de una tendencia media o lineal, o si no es estacionaria debido a una raíz unitaria. El KPSS test, es un caso especial de los test que prueban la estacionariedad mediante el estudio de raices unitarias pues, al contrario que la mayor parte de estos test, tiene como H_0 que la serie es estacionaria.

  H_0: Muestra estacionaria
  H_a: Muestra no estacionaria

  Lo que realiza realmente el KPSS test es descomponer la serie en la suma de una tendencia determinística, un camino aleatorio y un error tal que:

  $$x_t = \alpha t + u_t + e_t$$

  donde $u_t = u_{t-1} + a_t$ y los a_t son i.i.d $N(0, \sigma^2)$.
  La hipótesis nula de que X es estacionaria se da cuando $\sigma^2 = 0$. Para calcular el test estadístico se consideran los tres tipos de modelos al igual que para el test ADF:

  - $x_t = u_t + e_t$ donde se encuentra una tendencia determinística pero no desviación típica.
  - $x_t = \mu + u_t + e_t$ se encuentra desviación pero no tendencia
  - $x_t = \mu + \alpha_t + u_t + e_t$ se encuentra desviación y tendencia

  El estadístico de KPSS es

  $$KPPS = \sum_{i=1}^{T} S_t^2 / \hat{\sigma}_\varepsilon^2$$

  donde, $\hat{\sigma}_\varepsilon^2$ es el error estimado de la varianza de la regresión x y:

  $$S_t = \sum_{i=1}^{t} e_i, t = 1, \ldots, T \tag{3}$$

  con $e_t, t = 1, \ldots, T$ los valores residuales de la regresión de x..

En este proyecto el criterio que se sigue es del nivel de significación del 95 %, esto es, para aquellos valores resultantes del test de hipótesis menores que 0.05 se rechazará la hipótesis nula. Destacar que, en el caso de que la hipótesis nula no se rechace, esto no conlleva a decir que la hipótesis nula sea cierta, si no que no existen evidencias suficientes para rechazarla. Por tanto, nuestro objetivo es no tener evidencias suficientes para rechazar la hipótesis nula de independencia, rechazar la hipótesis nula de no estacionariedad y no tener evidencias suficientes para poder rechazar la hipótesis de estacionariedad en los test Box, ADF y KPSS respectivamente para poder concluir que nuestras muestras son dependientes y estacionarias.

3.4 Métodos de Normalidad

Para el estudio de la Gaussianidad de los datos se van a utilizar dos test[5]: Epps y Lobato y Velasco. Ambos consideran las siguientes hipótesis:

H_0: las variables X_i siguen una distribución normal para todo i en $\mathbb{Z}$
H_a: las variables no siguen una distribución normal

Sin embargo, puede que ninguno de los test rechace la hipótesis nula y que, por lo tanto, no se tengan evidencias suficientes para decir que los datos no siguen una distribución normal, ya que miran únicamente momentos de órdenes bajos. En concreto el test de Epps mira la media y varianza y el test de Lobato y Velasco la asimetría y curtosis.

A continuación, se van a introducir los distintos métodos que se han utilizado para la realización del test de la normalidad del presente trabajo.

[5]En el anexo se puede encuentrar el fichero test.m

3.4.1 Test de Epps

El test de Epps [12] compara $\phi_{X_t}(\lambda_i)$ con $\phi_{N(\mu,\sigma^2)}(\lambda_i)$, es decir, comprueba si la función característica de la marginal unidimensional de un proceso estrictamente estacionario coincide con la función característica de la distribución Gaussiana. Una ventaja a destacar de este test es que únicamente se necesita saber la media y la covarianza del proceso $(X_t)_{t\in\mathbb{Z}}$ tal que $\mathbb{E}(X) = \mu$ y $cov(X_0, X_r) = \sigma(r)$ con $r = 0, \pm1, \pm2\ldots$ donde la media de las funciones son componentes de la función característica empírica. La comparación directa de las dos funciones características no es posible en este caso, ya que no sabemos cuál es la función característica de nuestro proceso estocástico, por lo que Epps propone realizar la comparación de cada función en puntos determinados.

En primer lugar, se define Λ_N como un conjunto finito de valores reales positivos tal que: Λ_N:=$\{\lambda := (\lambda_1, \ldots, \lambda_N)^T \in \mathbb{R}_N^+ : \lambda_i \neq \lambda_j, i \neq j, i, j = 1, \ldots, N\}$ donde ^{T} denota la traspuesta. Ahora bien, se sabe que la función característica de una distribución normal con media $\nu \in \mathbb{R}$ y desviación típica $\rho > 0$ es:

$$\phi_{N(\nu,\rho^2)} = e^{i\nu t + \frac{\rho^2 t^2}{2}} \tag{4}$$

y sabemos que la fórmula de Euler relaciona esta ecuacion con senos y cosenos de la siguiente manera:

$$e^{ix} = cos(x) + isin(x)$$

donde $cos(x)$ es la parte real y $isin(x)$ la parte imaginaria. Con esto, Epps consideró oportuno formar un vector incluyendo como elementos la parte real e imaginaria de la función característica normal evaluada en determinados momentos λ_i de la manera siguiente:

$$g_{\nu,\rho}(\lambda) := (Re(\phi_{N(\nu,\rho^2)}(\lambda_1)), Im(\phi_{N(\nu,\rho^2)}(\lambda_1)), \ldots, Re(\phi_{N(\nu,\rho^2)}(\lambda_N)), Im(\phi_{N(\nu,\rho^2)}(\lambda_N)))^T.$$

Dada una muestra de observaciones igualmente espaciadas del proceso aleatorio X, $\{X_1, X_2, ..., X_n\}$, $n \in \mathbb{N}$ y sean $\lambda \in \Lambda_N$ y $\widehat{g}(\lambda)$ el vector columna 2N-dimensional compuesto por la parte real y compleja de la función carasterística de nuestro proceso X evaluada en determinados momentos λ_i tal que:

$$\widehat{g}(\lambda) : \frac{1}{n}\sum_{i=1}^{n}(\cos(\lambda_1 X_i), \sin(\lambda_1 X_i), \ldots, \cos(\lambda_N X_i), \sin(\lambda_N X_i))^T,$$

consideramos a $\widehat{g}(\lambda)$ como el estimador de $g_{\nu,\rho}(\lambda)$. A partir de ahora, para simplificar la notación denotaremos a $g_{\nu,\rho}(\lambda)$ como $g(\lambda)$. Entonces, fijándonos en $g(\lambda)$, obtenemos que la función de densidad de nuestro proceso es la denominada matriz de densidad espectral con frecuencia 0 siguiente:

$$f_X(0; (\mu_X, \gamma_X), \lambda) = (g(X_t, \lambda))_{t\in\mathbb{Z}} := ((\cos(\lambda_1 X_t), \sin(\lambda_1 X_t), \ldots, \cos(\lambda_N X_t), \sin(\lambda_N X_t)))_{t\in\mathbb{Z}}^T \tag{5}$$

Para la construcción del test estadístico, se utilizará el siguiente estimador de $f_X(0, (\mu_X, \gamma_X), \lambda)$:

$$\widehat{f}(0, \lambda) = (2\pi n)^{-1}\left(\sum_{t=1}^{n}\widehat{G}(X_t, \lambda) + 2\sum_{t=1}^{\lfloor n^{2/5}\rfloor}(1 - \frac{i}{\lfloor n^{2/5}\rfloor})\sum_{t=1}^{n-i}\widehat{G}(X_{t+i}, \lambda)\right) \tag{6}$$

donde $\widehat{G}(X_{t+i}, \lambda) = (g(X_t, \lambda) - \widehat{g}(\lambda))(g(X_{t+i}, \lambda) - \widehat{g}(\lambda))^T$ y $\lfloor\cdot\rfloor$ denota la parte entera. Observamos que este estimador está comparando la diferencia existente entre la función característica normal y la función característica estimada en determinados puntos. Además, se puede ver que tiene forma de covarianza, pues refleja cómo dos variables aleatorias, $g(X_t, \lambda)$ y $(g(X_{t+i}, \lambda)$, varían de forma conjunta con respecto a su media $\widehat{g}(\lambda)$.

El estimador $\widehat{f}(0, \lambda)$ es el utilizado en Epps [12] pero sustituyendo la fracción 2/5 por una constante general del intervalo (0, 1/2). Además, en Epps [12] se prueba que si $(X_t)_{t\in\mathbb{Z}}$ es Gaussiano, estacionario y satisface (7), entonces $\hat{f}(0, \lambda)$ converge casi seguro a $f_X(0, (\nu_X, \gamma_X), \lambda)$, es decir, converge a la función de densidad de una distribución normal.

Recordatorio 3.2 *Recordemos la convergencia casi segura de un conjunto de v.a's:*

Sea $X_1, X_2, \ldots$ una sucesión infinita de v.a's y sea X una v.a. Se dice que $X_1, X_2, \ldots$ converge a X casi seguro si $\mathbb{P}\{\omega \in \Omega : X_n(\omega) \to X(\omega)\} = 1$. Lo denotaremos como $X_n \to X$ c.s.

Entonces, Epps [12] presenta la siguiente forma general para calcular la Gaussianidad en un test de hipótesis:

Sea $G_n^+(\lambda)$ la inversa generalizada de $2\pi\hat{f}(0,\lambda)$, consideramos $Q_n(\nu,\rho,\lambda)$ la forma cuadrática siguiente:

$$Q_n(\nu,\rho,\lambda) := (\hat{g}(\lambda) - g(\lambda))^T G_n^+(\lambda)(\hat{g}(\lambda) - g(\lambda)). \tag{7}$$

Una **forma cuadrática** es una ecuación cuyos términos están compuestos por dos variables (iguales o distintas)[1]. Es decir, una forma cuadrática, no puede tener términos ni lineales ni constantes. Así bien, supongamos que A es una matriz simétrica $n \times n$, entonces, la forma cuadrática asociada a A es la función definida por:

$$f(x) = x^T Ax \text{ donde x es un vector columna}$$

Entonces, si $a_{i,j}$ son las entradas de A y $x_1, x_2, \ldots, x_n$ son las de x, la forma cuadrática se puede escribir de la forma:

$$f(x) = a_{11}x_1^2 + \ldots, a_{nn}x_n^2 + \sum_{i<j} a_{ij}x_i x_j$$

Notemos, además, que una forma cuadrática puede ser clasificada de la siguiente manera:

- Definida positiva: se dice que una función cuadrática es definida positiva si la imagen por medio de la forma cuadrática por medio de cualquier vector no nulo es estrictamente positivo, es decir si $f(\hat{x}) > 0$, para todo $\hat{x} \in \mathbb{R} - \{0\}$.
- Definida negativa: se dice que una función cuadrática es definida negativa si la imagen por medio de la forma cuadrática por medio de cualquier vector no nulo es estrictamente negativo, es decir, si $f(\hat{x}) < 0$, para todo $\hat{x} \in \mathbb{R} - \{0\}$.
- Semiefinida positiva: se dice que una función cuadrática es semidefinida positiva si la imagen por medio de la forma cuadrática por medio de cualquier vector es positiva o nula, es decir, si $f(\hat{x}) \geq 0$, para todo $\hat{x} \in \mathbb{R}$.
- Semidefinida negativa: se dice que una función cuadrática es semidefinida negativa si la imagen por medio de la forma cuadrática por medio de cualquier vector es negativo o nulo, es decir, si $f(\hat{x}) \leq 0$, para todo $\hat{x} \in \mathbb{R}$.
- Indefinida: se dice que una función cuadrática es indefinida si tenemos vectores cuya imagen por medio de la forma cuadrática es tanto positiva como negativa, es decir, si existe $\hat{x}, \hat{y} \in \mathbb{R}$ tal que $f(\hat{x}) < 0 \wedge f(\hat{y}) > 0$.

Existen varios métodos para clasificar una forma cuadrática, nosotros vamos a nombrar únicamente los dos que consideramos más comunes, estos son:

- *Método de valores propios*: Sean $\lambda_1, \lambda_2, \ldots, \lambda_n$ los valores propios de la matriz asociada obtenidos de calcular $|A - \lambda I| = 0$ donde I es la matriz identidad, obtendremos la siguiente clasificación:
  - Si $\lambda_1, \lambda_2, \ldots, \lambda_n > 0$, entonces es definida positiva.
  - Si $\lambda_1, \lambda_2, \ldots, \lambda_n < 0$, entonces es definida negativa.
  - Si $\lambda_1, \lambda_2, \ldots, \lambda_n \geq 0$, entonces es semidefinida positiva.
  - Si $\lambda_1, \lambda_2, \ldots, \lambda_n \leq 0$, entonces es semidefinida negativa.
  - Si $\exists i, j | \lambda_i < 0 \wedge \lambda_j > 0$, entonces es entonces es ndefinida.

- *Método de los menores principales*: Sean $H_1, H_2, \ldots, H_n$, los menores principales de la matriz A, que son los determinantes de una matriz cuadrada que contienen los elementos de la diagonal principal y son de distintos órdenes, entonces:
  - Si $H_1, H_2, \ldots, H_n > 0$, entonces es definida positiva.
  - Si $H_1, H_3, \ldots, H_{2k+1} < 0 \wedge H_2, H_4, \ldots, H_{2k+2}$, entonces es efinida negativa.
  - Si $H_1, H_2, \ldots, H_{n-1} > 0 \wedge H_n = 0$, entonces es semidefinida positiva.
  - Si $H_1, H_3, \ldots, H_{2k+1} < 0 \wedge H_2, H_4, \ldots, H_{2k} > 0 \wedge H_n = 0$, entonces es semidefinida negativa.
  - Si se da cualquier otro caso entonces es indefinida.

Volviendo al estudio de Epps, sea $\Theta \subset \mathbb{R} \times \mathbb{R}^+$ un conjunto acotado, es decir, un conjunto tal que todos sus puntos están a una distancia finita de cualquier punto dado, y abierto. Sea $\lambda \in \Lambda_N$, entonces enunciamos dos hipótesis:

Suposición A. El conjunto $\Theta_0(\lambda) := \{(\nu, \rho) \in \Theta : \phi_{\nu,\rho}(\lambda_i) = \phi_X(\lambda_i), i = 1, ..., N\}$ es denso en Θ, esto es que, entre dos números cualesquiera del conjunto siempre cabe otro de la misma naturaleza.

Además, este conjunto es discreto y va a contener a lo sumo un elemento a excepción de que los $\lambda_j' s$ sean racionales múltiplos de λ_1.

A continuación, se incluye una suposición sobre unas condiciones de regularización de las funciones involucradas en los puntos en $\Theta_0(\lambda)$. Esta suposición la utilizaremos en los resultados relacionados con el test de Epps.

Suposición B. Para cada $(\nu, \rho) \in \Theta_0(\lambda)$ tenemos que $f_X(0, (\nu, \rho), \lambda) = f_X(0, (\mu_X, \gamma_X), \lambda)$ y que

$$\left.\frac{\partial \phi_{x,y}(\lambda_i)}{\partial(x,y)}\right|_{(x,y)=(\nu,\rho)} = \left.\frac{\partial \phi_{x,y}(\lambda_i)}{\partial(x,y)}\right|_{(x,y)=(\mu_X,\gamma_X)}, \text{ i=1,...,N.}$$

Se introduce la notación que va a ser utilizada de ahora en adelante. Para un determinado proceso estacionario $X = (X_t)_{t\in\mathbb{Z}}$ se tiene que:

- $\mu_X := \mathbb{E}[X_0]$ es la media del proceso
- $\mu_X := \mathbb{E}[(X_0 - \mu_X)^k]$ con $k \in \mathbb{N}$ es el momento centrado de orden k
- $\gamma_X(t) := \mathbb{E}[(Y_0 - \mu_Y)(Y_t - \mu_Y)]$ con $t \in \mathbb{Z}$ es la autocovarianza de orden t

y para una muestra de observaciones equiespaciadas del proceso $X_1, \ldots, X_n$ con $n \in \mathbb{N}$, definimos los siguientes estimadores:

- $\hat{\mu}_X := \frac{1}{n}\sum_{i=1}^n X_i$ es la media muestral
- $\hat{\mu}_{X,k} := \frac{1}{n}\sum_{i=1}^n (X_i - \hat{\mu}_X)^k$ con $k \in \mathbb{N}$ es el momento centrado muestral de orden k
- $\hat{\gamma}_X(t) := \frac{1}{n}\sum_{i=1}^{n-|t|}(X_i - \hat{\mu}_X)(X_{i+|t|} - \hat{\mu}_X)$ con $|t| \leq n-1$ es la autocovarianza muestral de orden t

El teorema siguiente, probado en Epps [12], muestra la distribución a la que converge el estadístico involucrado en el test de Epps bajo la hipótesis nula.

Teorema 3.2 *Sea X un proceso Gaussiano estacionario que cumple la condición siguiente*

$$\sum_{t\in\mathbb{Z}} \mid t \mid^{\zeta} \mid \gamma_X(t) \mid < \infty \tag{8}$$

para algún $\zeta > 0$.
Sea $\Theta \subset \mathbb{R} \times \mathbb{R}^+$ *un conjunto abierto y acotado y* $\lambda \in \Lambda_N$ *tales que las suposiciones A y B se cumplen. Sea* (μ_n, γ_n) *el minimizador en* Θ *más cercano a* $(\widehat{\mu}_X, \widehat{\gamma}_X)$ *de la aplicación* $(\nu, \rho) \to Q_n(\nu, \rho, \gamma)$. *Asumamos además que* $f_X(0, (\mu_X, \gamma_X), \lambda)$ *es definida positiva. Entonces, para cada* $\lambda \in \Lambda_N$ *fijo,* $nQ_n(\mu_n, \gamma_n, \lambda)$ *converge en distribución a* χ^2_{2N-2}.

Recordemos que una sucesión de variables aleatorias *converge en distribución* si $\lim_{n\to\infty} F_n(x) = F(x)$ para todo $x \in \mathbb{R}$ con F continua y F_n y F la función de distribución acumulada de las variables aleatorias X_n y X respectivamente. Esto ocurre únicamente si $X_1, \ldots, X_n$ son variables aleatorias normales independientes de media cero y varianza uno.

Este test no es consistente contra las alternativas con marginales Gaussianas o, incluso, contra las distribuciones con las marginales no Gaussianas cuyas funciones caraterísticas toman los valores apropiados en los puntos seleccionados. En el teorema 4.1, vemos que este problema se mitiga al realizar la proyección aleatoria de todo el proceso, lo que hace que este test sea consistente contra todas las alternativas con marginales unidimensionales no Gaussianas.

3.4.2 Test de Lobato y Velasco

Lobato y Velasco [16] estudian los momentos centrados de las funciones características, en concreto, los momentos centrados considerados en el test de Gaussianidad son los órdenes 3 y 4 que corresponden, respectivamente, a la asimetría y curtosis del proceso. Como apunte, mencionar que la curtosis es una medida de forma, es decir, mide cuanta cúpula o cuan achatada está una curva o distribución. Cuanto más alto sea el valor de la curtosis, más puntiaguda será la curva. Por otra parte recordemos la definición de momento muestral y momento centrado de orden k:

Definición 3.2 *Sea una muestra de v.a $X_1, X_2, \ldots, X_n$ y $k \in \mathbb{N}$, el momento muestral de orden k es:*

$$m_k := \tfrac{1}{n} \sum_{i=1}^{n} X_i^k$$

Mientras que el momentro centrado (con respecto a la media) de orden k es:

$$M_k := \tfrac{1}{n} \sum_{i=1}^{n} (X_i - m_1)^k$$

donde m_1 es la media muestral y M_2 la varianza muestral.

Entonces, el estadístico propuesto por Lobato y Velasco [16] para probar la Gaussianidad del proceso X es:

$$\tilde{G}_X := \frac{n\widehat{\mu}_{X,3}^2}{6\tilde{F}_3} + \frac{n(\widehat{\mu}_{X,4} - 3\widehat{\mu}_{X,2}^2)^2}{24\tilde{F}_4} \tag{9}$$

donde

$$\tilde{F}_k := 2\sum_{t=1}^{n-1} \widehat{\gamma}_X(t)(\widehat{\gamma}_X(t) + \widehat{\gamma}_X(n-t))^{k-1} + \widehat{\gamma}_X^k$$

es el estimador de:

$$F_k := \sum_{t=-\infty}^{\infty} \gamma_X(t)^k$$

El teorema siguiente muestra el comportamiento del test de Lobato y Velasco.

Teorema 3.3 *(Lobato and Velasco,2004)*
Sea $X = (X_t)_{t\in\mathbb{Z}}$ un proceso estacionario ergódico.

- *Si X es Gaussiano y cumple $\sum_{t=0}^{\infty} \mid \gamma_X(t) \mid < \infty$, entonces $\tilde{G}_X \to \chi_2^2$ converge en distribución.*
- *Si $\mu_{X,3} \neq 0$ o $\mu_{X,4} \neq 3\mu_{X,2}^2$, $\tilde{G}_X$ diverge a infinito cuando cumple las siguientes condiciones:*
  - *Si $\mathbb{E}[X_t^{16}] < \infty$*
  - *$\sum_{t_1=\infty}^{\infty} \cdots \sum_{t_{q-1=\infty}}^{\infty} \mid k_q(t_1, \ldots, t_{q-1}) \mid < \infty$, para q=2,...,16, donde $k_q(t_1, \ldots, t_{q-1})$ denota el q-ésimo orden cumulativo de $X_1, X_{1+t_1}, \ldots, X_{1+t_{q-1}}$*

- $\sum_{t=1}^{\infty}[\mathbb{E} \mid (\mathbb{E}(X_0-\mu)^k \mid \mathscr{F}_{-t}) - \mu_k \mid^2]^{1/2} < \infty$, *para k=3,4, donde* $\mathscr{F}_{-t}$ *denota el campo* σ *generado por* $X_j, j \leq -t$
- $\mathbb{E}[(X_0-\mu)^k - \mu_k]^2 + 2\sum_{t_1=\infty}^{\infty}\mathbb{E}([(X_0-\mu)^k-\mu_k][(X_t-\mu)^k-\mu_k]) > 0, k = 3, 4.$

Como se muestra en el teorema anterior, el test de hipótesis no es consistente ya que este test solo comprueba si la curtosis y asimetría de la marginal coincide con las de la distribución Gaussiana. Una vez más, este problema se solucionará aplicando unas pequeñas modificaciones al estadístico que se verá en la Sección 4.

3.5 Test múltiple y False Discovery Rate

Cuando realizamos distintos test con modelos matemáticos, en los resultados, observamos cierto error. En este estudio, como ya se ha dicho anteriormente, el método que se utiliza es el de los test de hipótesis múltiple y, los posibles valores obtenidos que podrán alterar nuestra conclusión, serán tanto los errores de tipo I (falsos positivos) como los errores de tipo II (falsos negativos). Los falsos positivos surgen cuando la hipótesis nula es rechazada y, sin embargo, se sabe que es verdadera. Por otra parte, el falso negativo se da cuando la hipótesis nula es rechazada mientras que debería ser aceptada. De manera general, cuando las muestras vienen de la misma distribución los p-valores están uniformemente distribuidos. Sin embargo, cuando las muestras vienen de distribuciones distintas, los p-valores están muy distantes de su media matemática y próximos a cero.

En un principio, para dar solución a los falsos positivos, existía unicamente la tasa de error del FWER (Family Wise Error). Sin embargo, esta tasa de error está basada en controlar la probabilidad de rechazar erróneamente incluso una de las hipótesis nulas verdaderas y no cuántas hipótesis nulas se pueden rechazar. De la necesidad de potenciar y afinar el problema de los test de hipótesis múltiples nació el False Discovery Rate (FDR). El FDR es la proporción esperada de las hipótesis rechazadas incorrectamente durante los k test realizados. Así, el FDR se utiliza para limitar la tasa de error en los test estadísticos. Siendo exactos, el FDR en sí no es un método para dar solución a los falsos positivos, pero el término se utiliza intercambiablemente con los métodos. En particular, el FDR se utiliza en el método de Benjamini - Hochberg [3] y Benjamini-Yekutieli [23]. El procedimiento de control del FDR en los test múltiples, es un procedimiento escalonado que involucra a un conjunto lineal de constantes en la escala de los p-valores. El FDR está relacionado al test global de las intersecciones de hipótesis, que está definido en terminos del mismo conjunto de constantes: rechazar la intersección única de hipótesis si existe un i tal que $p_{(i)} \leq \frac{i}{m}\alpha$. La distinción entre un test global y un procedimiento de test múltiple es importante. Si la única hipótesis de intersección es rechazada por un test global, uno no puede apuntar a las hipótesis por separado ya que pueden ser falsas. Mientras que unas hipótesis son verdaderas y otras son falsas, el test global no controla necesariamente el FWER al nivel deseado, por lo que no debería tratarse como un procedimiento de tests múltiples.

El interés en la actuación del test global cuando los test estadísticos son dependientes empezó con Simes (1986), quien investigó si el procedimiento era conservativo bajo alguna estructura dependiente utilizando simulaciones. El test es conservativo para estadísticos con dependencia positiva. Procedamos entonces a comparar el procedimiento del multiple testing de Bonferroni con el FDR de Hochberg y el de Yekutieli:

3.5.1 Procedimiento de Bonferroni

Sean los test de hipótesis $H_1, H_2, ..., H_m$, donde m es el número de test realizados y sus correspondientes p-valores son $p_1, p_2, ..., p_m$. Sean $P_{(1)} \leq P_{(2)} \leq ... \leq P_{(m)}$ donde H_i es la hipótesis nula correspondiente a P_i. Se define el procedimiento de Bonferroni:

Sea k el mayor i para el que

$$P_{(i)} \leq \frac{i}{m}\alpha$$

donde α es el nivel de significación. Entonces, rechazamos todos los $H_{(i)}, i = 1, ..., k$. Se va a trabajar con el p-valor definido a continuación:

$$p_0 = m \min_{i=1,...,m} \{P_{(i)}/i\} \tag{10}$$

En el caso de que todas las hipótesis probadas sean verdaderas, es decir, cuando $m_0 = m$, este teorema se reduce a la prueba global de Simes de la hipótesis de intersección probada primero por Seeger (1968) y luego independientemente por Simes (1986). Sin embargo, cuando $m_o < m$ el procedimiento no controla el FWER. Para lograr el control de FWER, Hochberg (1988) construyó un procedimiento a partir de la prueba global, que tiene la misma estructura escalonada pero cada $P_{(i)}$ se compara con $m - p_1$ en lugar de i. Las constantes para los dos procedimientos son las mismas en $i = 1$ e $i = m$ pero en otras partes las constantes de control FDR son más grandes. Enunciemos entonces el procedimiento mencionado.

3.5.2 Procedimiento de Benjamini - Hochberg

La tasa de falsos descubrimientos (FDR), sugerida por Benjamini y Hochberg (1995)[3] es un punto de vista nuevo y diferente sobre cómo podrían considerarse los errores en los test de hipótesis múltiples ya que, habitualmente se acostumbra a preguntar si se ha realizado algún error en lugar de preguntarse cuántos errores se han realizado. Se debe de saber que, cuando todas las hipótesis nulas son verdaderas, el control del FDR es equivalente al control del FWER, mientras que, cuando muchas hipótesis nulas son rechazadas el control es más pequeño. Por lo tanto, somos capaces de soportar más errores cuando se rechazan muchas hipótesis, pero soportaremos menos cuantas menos hipótesis se rechacen. Deseamos entonces, hacer tantos descubrimientos como sea posible sujetos al control del FDR.

Procedimiento de Hochberg (1995):
Sean los test de hipótesis $H_1, H_2, ..., H_m$, donde m es el número de test realizados y sus correspondientes p-valores son $p_1, p_2, ..., p_m$. Sean $P_{(1)} \leq P_{(2)} \leq ... \leq P_{(m)}$ donde H_i corresponde a la hipótesis nula del p-valor p_i. El procedemiento de Hochberg es el definido a continuación:

Sea k el mayor i para el que

$$P_{(i)} \leq \frac{i}{m + 1 - i}\alpha$$

entonces, rechazamos todos los $H_{(i)}$, $i = 1, ..., k$.

De acuerdo al procedimiento de Hochberg y, dado que nosotros trabajamos con una sola hipótesis, entonces rechazaremos a un nivel de significación α cuando se cumpla la siguiente desigualdad:

$$\frac{m + 1 - i}{i} P_{(i)} \leq \alpha$$

Entonces, al estar trabajando con una única hipótesis, nos bastará con rechazar al menos una hipótesis para poder concluir que nuestro proceso no es Gaussiano. En la teoría de Hochberg no aparece ningún p-valor, pero nosotros se sugiere el p-valor siguiente a partir del cual se va a rechazar la hipótesis nula:

$$p_0 = \min_{i=1,...,m} \left\{(m + 1 - i)P_{(i)}/i\right\} \tag{11}$$

Es de notar la relación existente entre el procedimiento de Hochberg y el procedimiento del control del FDR (Bonferroni). Ambos son procedimientos que empiezan comparando $P_{(m)}$ con α y, si el p-valor es más pequeño ($P_{(m)} \leq \alpha$), entonces, se rechazan todas las hipótesis. Si $P_{(m)} \geq \alpha$, se procede con un p-valor hasta que uno satisfaga la condición. Si el procedimiento termina sin haber encontrado un p-valor menor que α, realiza la comparación siguiente: $P_{(1)}$ con α/m. En este sentido, los dos procedimientos siguen el mismo patrón, sin embargo, en Bonferroni cada $P_{(i)}$ se compara con $\{1 - (i - 1)/m\}\alpha$, mientras que en el procedimiento de Hochberg se compara con $\{1/(m + 1 - i)\}\alpha$.

3.5.3 Procedimiento de Benjamini y Yekutieli

En Benjamini y Hochberg, se comprueba que el FDR controla estos errores mejor que el Family Wise Error. Sin embargo, Benjamini y Yekutieli [23] [4], probaron que el mismo procedimiento también controla el FDR cuando el test estadístico tiene una dependencia de regresión positiva en cada uno de los test estadísticos correspondientes a la hipótesis nula. Esta condición de dependencia positiva suele ser suficiente para cubrir muchos problemas de interes práctico.

Procedimiento de Benjamini y Yekutieli

Teorema 3.4 *(Benjamini and Yekutieli(2001)). Asumamos que aplicamos k tests estadísticos para comprobar la misma hipótesis nula y que los p-valores que obtenemos son* $p_{(1)}, ..., p_{(m)}$*, donde* $p_{(1)} \leq ... \leq p_{(m)}$*.*

Sea $\alpha \in (0,1)$*. El FDR del test que rechaza la hipótesis nula, si el conjunto* $\left\{i : p_{(i)} \leq \frac{i\alpha}{m\sum_{j=1}^{m} j^{-1}}\right\}$ *no es vacío es, a lo sumo,* α*.*

Notemos que si $\sum_{j=1}^{m} j^{-1} = 1$ tendríamos $\frac{i\alpha}{m}$, lo que sería el procedimiento de Bonferroni.

Así, asumiendo que todas las hipótesis son iguales y de acuerdo al teorema anterior, sugerimos que

$$p_0 := m\sum_{j=1}^{m} j^{-1} \min_{i=1,...,m} p_{(i)}/i \tag{12}$$

sea el valor para el que vamos a poder rechazar la hipótesis nula a cualquier nivel $\alpha \geq p_0$ y, entonces, tomar p_0 como el p-valor resultante del procedimiento.

Ejemplo 2. Se obtiene una muestra de la misma distribución de 10 p-valores con nivel de significación α =0.05 obtenidos al realizar un test de hipótesis múltiple:

p-valores:	0.2040	0.0021	2.0849e-05	0.2088	0.2095	0.2099	0.2126	2.1269e-06	0.2137

En primer lugar, ordenamos los p-valores de menor a mayor:

p-valores:	2.1269e-06	2.0849e-05	0.0021	0.2040	0.2088	0.2095	0.2099	0.2126	0.2137
	P(1)	P(2)	P(3)	P(4)	P(5)	P(6)	P(7)	P(8)	P(9)

- Calculamos el FDR mediante el método de **Bonferroni** [3]:

  Calculamos el p-valor definido en (10) y veamos si cumple que $p_0 \leq 0{,}05$:

  $P_0 = 9\,\text{mín}\{P(i)/i\}$

  $P_0 = 9\,\text{mín}\{\frac{P(1)}{1}, \frac{P(2)}{2}, \frac{P(3)}{3}, \frac{P(4)}{4}, \frac{P(5)}{5}, \frac{P(6)}{6}, \frac{P(7)}{7}, \frac{P(8)}{8}, \frac{P(9)}{9}\}$

  $P_0 = 9(2{,}1269e-06) = 1{,}91421e-05$

  Entonces, como el p_0 calculado es menor que 0.05, podemos concluir que el proceso del que obtenido se ha obtenido la muestra no sigue una distribución normal mediante el ajuste de Bonferroni.

- Ahora, calculamos el FDR mediante el método de **Hochberg** [3]:

  Realizamos el cálculo del p-valor definido en (11) tal que:

  $P_0 = \text{mín}\{(9+1-i)P(i)/i\}$

  $P_0 = \text{mín}\left\{9\frac{P(1)}{1}, 8\frac{P(2)}{2}, 7\frac{P(3)}{3}, 6\frac{P(4)}{4}, 5\frac{P(5)}{5}, 4\frac{P(6)}{6}, 3\frac{P(7)}{7}, 2\frac{P(8)}{8}, 1\frac{P(9)}{9}\right\}$

$$P_0 = 1{,}91421e-05$$

Hemos obtenido que p_0 es menor que 0.05, por lo que podemos concluir que el proceso del que obtenido se ha obtenido la muestra no sigue una distribución normal mediante el ajuste de Hochberg.

- A continuación, utilizamos ahora el método de **Yekutieli** [4]:

  Realizamos el cálculo (12):

$$p_0 := 9\sum_{j=1}^{9}\frac{1}{j}min\{\frac{P(1)}{1},\frac{P(2)}{2},\frac{P(3)}{3},\frac{P(4)}{4},\frac{P(5)}{5},\frac{P(6)}{6},\frac{P(7)}{7},\frac{P(8)}{8},\frac{P(9)}{9}\}$$

$$p_0 := 9\sum_{j=1}^{9}\frac{1}{j}(2{,}1269e-06)$$

$$p_0 := 9(2{,}8290)(2{,}1269e-06)$$

$$p_0 := 5{,}4153e-05.$$

  Dado que $p_0 \leq 0{,}05$, podemos rechazar a cualquier nivel $\alpha > p_0$. Entonces concluimos que el proceso del que se ha obtenido la muestra no sigue una distribución normal.

4 Procedimiento en la práctica

En (1) la hipótesis nula se cumple si y solo si $(X_1,\ldots,X_t)^T$ es un vector Gaussiano para todo $t \in \mathbb{N}$. Debido a la estacionariedad de X, esto es equivalente a que $(X_t)_{t\leq 0}$ sea Gaussiano y, por ello, también es equivalente a la Gaussianidad del proceso $X^{(t)} := (X_j)_{j\leq t}$ para cualquier $t \in \mathbb{Z}$. Dado $t \in \mathbb{Z}$, queremos usar el Teorema 3.1 para comprobar si $X^{(t)}$ es Gaussiano. Por tanto, el procedimiento que vamos a realizar para hallar si nuestro proceso sigue una distribución normal es el siguiente:

1. Incluir $X^{(t)}$ en un espacio de Hilbert apropiado, $\mathbb{H}$.
2. Seleccionar un vector aleatorio $h \in \mathbb{H}$ usando la función disipativa.
3. Calcular $\langle X^{(t)}, h\rangle$

Entonces tendremos que $X^{(t)}$ es Gaussiano, si y solo si, el valor real $\langle X^{(t)}, h\rangle$ es Gaussiano.

4.1 Determinación del espacio de Hilbert

Consideremos que nuestro proceso $X^{(t)}$ pertenece al espacio de sucesiones de Hilbert $\mathbb{H}$ siguiente: $l^2 = \left\{(x_n)_{n\in\mathbb{N}} : \sum_{n\in\mathbb{N}} x_n^2 a_n < \infty\right\}$, con $a_0 := 1$ y $a_n := \frac{1}{n^2}, n \geq 1$, dotado del producto escalar:

$$\langle x, y\rangle = \sum_{n\in\mathbb{N}} x_n y_n a_n, \text{ donde } x = (x_n)_{n\in\mathbb{N}} \text{ y } y = (t_n)_{n\in\mathbb{N}}.$$

Si el proceso es estacionario y la varianza de X_t es finita, entonces $\mathbb{E}[\sum_{n\in\mathbb{N}} X_{t-n}^2 a_n]$ también es finita. Esto implica que, casi seguro, $X^{(t)} \in l^2$.

4.2 Determinación del vector aleatorio $h \in \mathbb{H}$

Se indica a continuación el vector aleatorio $h \in \mathbb{H}$. Se necesita una distribución disipativa en l^2 para poderla utilizar en la selección del vector mediante el que se proyectarán los datos. Para hacer esto, va a ser utilizada la llamada distribución de Dirichlet *(Pitman,2006)* y la construiremos utilizando el método iterativo:

Sea $(\beta_n)_n \in \mathbb{N}$ independiente e igualmente distribuido con la distribución *beta* de parámetros $\alpha_1, \alpha_2 > 0$. Consideremos la distribución de un punto aleatorio del espacio l^2 obteniéndolo de la siguiente manera:
• $l_0 \in [0,1]$ es calculada con la distibución $\beta(\alpha_1, \alpha_2)$ y,
• para $n \geq 1$, $l_n \in [0, 1 - \sum_{i=0}^{n-1} l_i]$ es calculada multiplicando una variable aleatoria independiente $\beta(\alpha_1, \alpha_2)$ por 1-$\sum_{i=0}^{n-1} l_i$.

Definimos entonces $H_n = (l_n/a_n)^{1/2}$ para $n \geq 0$ y fijamos $H = (H_0, H_1, \ldots)^T$. Se puede comprobar fácilmente que la distribución del proceso estocástico $H = (H_n)_{n\in\mathbb{N}}$ es disipativa (ver definición 3.1) y que $h = (h_i)_i \in \mathbb{N}$ es una realización de H. El único punto que queda es probar que los elementos generados por esta distribución pertenecen a l^2. Lo vemos a continuación:

Proposición 4.1 *Sea $H = (H_n)_{n\geq 0}$ un proceso estocástico construido como se ha descrito anteriormente. Entonces $\|H\| = 1$ casi seguro.*

Entonces nuestra $h = (h_i)_{i\in\mathbb{N}}$ es una realización de H con $h = h_0, ...h_{m-1}$ y h_m tal que $\|h\| = 1$.

4.3 Construcción de la proyección y resultados en los que nos apoyamos

Tenemos entonces un proceso estacionario $X = (X_t)_{t\in\mathbb{Z}}$ y un vector aleatorio $h := (h_t)_{t\in\mathbb{N}}$ de H con X y H independientes entre sí. Construímos un nuevo proceso $Y = (Y_t^h)_{t\in\mathbb{Z}}$ dado por las proyecciones de $X = (X^{(t)})_{t\in\mathbb{Z}}$ en el espacio unidimensional generado por h tal que:

$$Y_t := \langle X, h\rangle = \sum_{i=0}^{\infty} h_i X_{t-i} a_i, t \in \mathbb{Z} \tag{13}$$

El Teorema 3.1 citado en la Sección 3 implica que, si X no es Gaussiano, entonces la h que hemos elegido hace que Y_t sea no Gaussiana. En otras palabras, si X no es Gaussiana, entonces la marginal unidimensional de Y no es gaussiana para casi ninguna h.

Por otra parte, denotemos $\gamma_{Y|h}(t) := \mathbb{E}[(Y_0 - \mu_{Y|h})(Y_t - \mu_{Y|h}) \mid h]$ la autocovarianza condicionada, donde $\mu_{Y|h} := \mathbb{E}[Y_0 \mid h]$ es la esperanza condicionada de Y_0 dada h.

Recordemos que la esperanza condicionada de Y dada $H = h$ es:

$$\mathbb{E}[Y \mid H = h] = \sum_Y y_i \mathbb{P}_{Y|H}(y_i \mid h) \text{ donde } \mathbb{P}_{Y|H}(y_i \mid h) = \frac{\mathbb{P}_{YH}(y_i, h_j)}{\mathbb{P}_H(h_j)}$$

La siguiente proposición muestra que el proceso proyectado mantiene las propiedades del proceso original.

Proposición 4.2 *Sea $(X_t)_{t\in\mathbb{Z}}$ un proceso ergódico y estacionario tal que $\sum_{t=0}^{\infty} t^\zeta \mid \gamma_X(t) \mid < \infty$, con $\zeta \geq 0$. Entonces, con las condiciones de h, el proceso $(Y_t)_{t\in\mathbb{Z}}$ definido en (1) es ergódico y estacionario. Además, $\mathbb{E}[\mid Y_0 \mid\mid h]$ y $\sum_{t=0}^{\infty} t^\zeta \mid \gamma_{Y|h}(t) \mid$ son finitos.*

Donde un proceso es ergódico si todos sus parámetros estadísticos se pueden determinar con una única realización del proceso. Esto implica que los distintos parámetros estadísticos se pueden expresar como medias temporales. Por lo que se puede concluir que un proceso estacionario es ergódico para la estimación de la media μ si las medias temporales coinciden con las medias estadísticas. Como la media no depende del tiempo, tendremos que Y es ergódico si $Var(\overline{Y}) \to \mu$ cuando $n \to \infty$.

Entonces podemos concluir que el proceso Y_t es también estacionario y, por tanto, podemos emplear los test mencionados en la Sección 3.5 para probar la Gaussianidad de las marginales de Y_t ya que han sido diseñados, precisamente, para probar la Gaussianidad de una marginal unidimensional. Con esto, de acuerdo al razonamiento precedente, estamos probando la completa Gaussianidad de X. En el caso finito dimensional, la distribución disipativa y la distribución absolutamente continua con respecto a la medida de Lebesgue coinciden. Así, las distribuciones disipativas se pueden considerar como una generalización de las distribuciones absolutamnete continuas del caso dimensional infinito en donde no hay medida que haga el rol de la medida de Lebesgue. Debería destacarse que todos las distribuciones no degeneradas Gaussianas son disipativas.

El Lema 4.1 descrito más abajo muestra que si los puntos involucrados en el test de Epps son seleccionados aleatoriamente, entonces la consistencia del estimador de la matriz espectral de densidad en 0 es mayor. Para estabilizar el resultado, necesitamos algunos resultados preliminares, que concluyen en un corolario que muestra que el test de Epps se comporta propiamente cuando se aplica al proceso Y.

Denotamos como $k_{lmno}(q, r, q+r; \lambda)$ el cuarto orden cumulativo de $Z_{0,l}, Z_{q,m}, Z_{r,n}$ y $Z_{q+r,o}$, donde, por ejemplo, $Z_{q,m}$ es la m-ésima componente del vector $g(Y_q, \lambda) - g_{\mu_Y, \gamma_Y}(\lambda)$.

Lema 4.1 *Sea $\lambda \in \Lambda_N$ y sea Y un proceso estacionario tal que*

$$\sup_{-\infty<q<\infty} \sum_{r=-\infty}^{\infty} \mid k_{lmno}(q, r, q+r; \lambda) \mid < \infty \tag{14}$$

para cada $l, m, n, o \in \{1, \ldots, N\}$. Entonces $\widehat{f}(0, \lambda) \to f_Y(0, (\mu_Y, \mu_Y), \lambda)$ casi seguro.

Recordemos que el supremo es la más pequeña de todas las cotas superiores y en caso de existir es único. Además, el supremo pertenece al conjunto si coincide con el máximo.

Lema 4.2 *Si $\lambda = (\lambda_1, \ldots, \lambda_N)^T \in \Lambda_N (N > 1)$ es hallada de tal manera que λ_1 y λ_2 son independientes e idénticamente distribuidas y tienen densidad, entonces la Suposición A del Teorema 3.2 se cumple casi seguro.*

El siguiente corolario deriva directamente del Teorema 3.2 y del lema anterior.

Corolario 4.1 *Sea $(Y_t)_{t\in\mathbb{Z}}$ un proceso Gaussiano estacionario tal que cumple (7) y construyamos λ como en el Lema 4.2. Sea (μ_n, γ_n) el minimizador en Θ más cercano a $(\widehat{\mu}, \widehat{\gamma})$ de la aplicación $(\nu, \rho) \to Q_n(\nu, \rho, \lambda)$. Si $f_Y(0, (\mu_Y, \gamma_Y, \lambda))$ es definida positiva, entonces $nQ_n(\mu_n, \gamma_n, \lambda)$ converge en distribución a χ^2_{2N-2}.*

El resultado siguiente proporciona las condiciones que permiten aplicar el test de Epps al proceso proyectado. Así, modificamos el E-test para seleccionar aleatoriamente los valores de λ. Esto mejora la consistencia del procedimiento inicial que ahora es capaz de detectar (con una muestra lo suficientemente grande) cada alternativa no Gaussiana que satisface las suposiciones.

Teorema 4.1 *Sea X un proceso estacionario que cumple $\sum_{t=0}^{\infty} \mid t \mid^{\zeta} \mid \gamma_X(t) \mid < \infty$ para algún $\zeta > 0$. Construimos λ como en el Lemma 4.2 y h independientemente de λ utilizando P_H. Asumimos que, con las condiciones en h, la Y definida en (13) satisface (14). Más allá de eso, asumamos también que los módulos de la función característica de su marginal unidimensional es analítica* [6] *y que $f_{Y|h}(0, (\mu_{Y|h}, \gamma_{Y|h}), \lambda)$ existe y es definida positiva para casi cualquier h.*

Sea $Q_n(\cdot, \cdot, \lambda)$ la forma cuadrática definida en la Sección 3.4.1 aplicado a Y y (μ_n, γ_n) el minimizador de Θ más cercano a $(\widehat{\mu}_{Y|h}, \widehat{\gamma}_{Y|h}), \lambda)$ de $Q_n(\cdot, \cdot, \lambda)$. Sea además,

$B := \{(\lambda, h) : nQ_n(\mu_n, \gamma_n, \lambda) \to_d$ una distribución no degenerativa$\}$.

Entonces, X es Gaussiano si y solo si $(P_\lambda \bigotimes P_H)[B] > 0$.

[6] Es la suma de una serie de potencias complejas indefindamente derivable en función de dicha variable en su dominio de convergencia

El corolario siguiente muestra que la consistencia del test de Epps mejora si los puntos involucrados son elegidos aleatoriamente.

Corolario 4.2 *Sea X un proceso ergódico y estacionario. Asumimos que el módulo de la función característica de su marginal unidimensional es analítica. Es más, asumimos también que (7) se cumple. Tomamos λ como en el Lema 4.2 y $Q_n(\cdot,\cdot,\lambda)$ como en la Sección 3.4.1. Sea (μ_n,γ_n) el minimizador de Θ más cecano a $(\widehat{\mu}_X,\widehat{\gamma}_X)$ de $Q_n(\cdot,\cdot,\lambda)$. Sea*

$C := \{\lambda : nQ_n(\mu_n,\gamma_n,\lambda) \to_d$ una distribución no degenerativa$\}$

Si asumimos que $f_x(0,(\mu_X,\gamma_X),\lambda)$ existe y es definida positiva, entonces, X es Gaussiano si y solo si $P_\lambda(C) > 0$

A continuación, enunciamos el siguiente corolario que establece un tipo de ley de cero a uno para reforzar las afirmaciones del Teorema 4.1 y Corolario 4.2.

Corolario 4.3 *Bajo las suposiciones del Teorema 4.1, $(P_\lambda \bigotimes P_H)[B] \in \{0,1\}$ y , X es Gaussiano si y solo si $(P_\lambda \bigotimes P_H)[B] = 1$.*

Análogamente, bajo la suposición del Corolario 4.2, $P_\lambda(C) \in \{0,1\}$ y X es Gaussiano, si y solo si, $P_\lambda(C) = 1$.

Observaciones 4.1 *Del Teorema 3.2 tenemos que el Teorema 4.1 y los Corolarios 4.2 y 4.3 se mantienen ciertos si sustituimos en la definición de conjuntos B y C "distribución no degenerativa"por "Distribución chi-cuadrado con 2(N-1) grados de libertad"; esto permite que el test sea construido basado en la distribución asintótica de $nQ_n(\mu_n,\gamma_n,\lambda)$.*

Para terminar esta sección, enunciamos un resultado que muestra la aplicabilidad del LV-Test al proceso proyectado bajo diferentes suposiciones de las usadas en Lobato y Velasco (2004). Para tal fin, reemplazamos el estadístico $\widehat{G}_Y$ por

$G_y = n\widehat{\mu}_3^2/(6 \mid \widehat{F}_3 \mid) + n(\widehat{\mu}_4 - 3\widehat{\mu}_2^2)^2/(24 \mid \widehat{F}_4 \mid)$,

con

$\widehat{F}_k = 2\sum_{t=1}^{\tau_n} \widehat{\gamma}(t)(\widehat{\gamma}(t) + \widehat{\gamma}(\tau_n + 1 - t))^{k-1} + \widehat{\gamma}^k, \tau_n < cn^{\beta_0}, 0 < \beta_0 < 0{,}5$ y $c > 0$.

Así, las diferencias entre G_Y y $\widehat{G}_Y$ son los valores absolutos en el denominador y el número de términos involucrados en $\widehat{F}_k$.

Teorema 4.2 *Sea X un proceso ergódico y estacionario que satisface $\sum_{t=0}^{\infty} \mid \gamma_X(t) \mid < \infty$. Entonces,*
1. Si X es un proceso Gaussiano, entonces $G_Y \to_d \chi_2^2$.
2. Asumamos que $X_t - \mu_X = \sum_{i=1}^{\infty} k(i)\epsilon_{t-i}$ con $\sum_{i=1}^{\infty} |k(i)| < \infty$,$\sum_{i=1}^{\infty} ik(i) < \infty$, y (ϵ_t) son variables aleatorias independientes e igualmente distribuidas con $\mathbb{E}[\epsilon_n] = 0$, y $\mathbb{E}[X_0^4] < \infty$. Así con las condiciones en h, G_Y diverge casi seguro al infinito cuando $\mu_3 \neq 0$ or $\mu_4 \neq 3\mu_2^2$.

Aplicando directamente el Teorema 4.2 al proceso X , obtenemos el siguiente corolario.

Corolario 4.4 *Bajo las suposiciones del Teorema 4.2, tenemos que si X es un proceso Gaussiano, entonces $G_X \to_d \chi_2^2$. Es más, si la suposición en el Punto 2 del teorema se sostiene, entonces bajo las condiciones en h, G_X diverge casi seguro al infinito cuando $\mu_{X,3} \neq 0$ o $\mu_{X,4} \neq 3\mu_{X,2}^2$.*

5 Resultados

En el presente apartado se van a presentar los resultados obtenidos al realizar el procedimiento expuesto a lo largo del proyecto.

Los datos de las alturas de las olas del mar que van a ser estudiados han sido obtenidos de *Coastal data information program* [11]. Existen distintas estaciones en las que se han colocado las boyas que están equipadas para tener la capacidad de comunicación satelital Iridum y que permiten la medición de la altura del mar. Estas boyas, realizan medidas en tres dimensiones, por lo que en nuestros conjuntos de datos tendremos 3 coordenadas a estudiar: x, y, z las cuales vamos a denotar como N, W y V respectivamente y cuya unidad de medida son centímetros (cm).

A continuación, enunciamos la nomeclatura utilizada para denominar cada una de las estaciones de las que se ha extraído los datos (ver Cuadro 1):

ESTACIÓN	**Coordenada X**	**Coordenada Y**	**Coordenada Z**
Santa Mónica Bay 028	28N	28W	28V
Point Reyes 029	29N	29W	29V
Grays Harbor 036	36N	36W	36V
Cape Mendocino 094	94N	94W	94V
Rincón 181	181N	181W	181V
Santa Lucía Escarpment 222	222N	222W	222V
Wallops Island 224	224N	224W	224V
Kaneohe Bay 225	225N	225W	225V
Pulley Ridge 226	226N	226W	226V
Santa Barbara 234	234N	234W	234V
Duck FRF 433	433N	433W	433V

Cuadro 1: Nomeclatura de las estaciones de las que se han obtenido los datos

Como hemos comentado, cada estación viene dada por tres series temporales y cada una de las series temporales está en $\mathbb{R}$. Se ha decidido realizar de esta manera el análisis debido a que los test que se han propuesto están creados para realizar el estudio de series temporales en $\mathbb{R}$, es cierto que también se podría haber realizado el análisis en $\mathbb{R}^3$ y hacer el análisis conjunto proyectando de $\mathbb{R}^3$ en $\mathbb{R}$, pero he considerado que hacer cada una de las series por separado sería realizar un estudio más detallado.

De ahora en adelante se van a exponer los resultados obtenidos realizando el proceso definido durante todo el proyecto. En primer lugar, se presentarán los resultados obtenidos al estudiar la estacionariedad ya que es condicion necesaria para poder seguir con el análisis. A continuación, se representan los resultados obtenidos al realizar los test de Normalidad sin realizar ninguna proyección. En este caso, si obtenemos algún rechazo de la hipótesis nula, no seguiremos realizando el estudio para esas series temporales pues hemos obtenido un resultado consistente. Por otra parte, para aquellos conjuntos de datos para los que no se ha conseguido el rechazo de la hipótesis de Gaussianidad, procederemos a representar los datos obtenidos al realizar el testing múltiple junto con el FDR correspondiente.

5.1 Resultados Dependencia y Estacionariedad

El análisis de la estacionariedad de los procesos estocásticos es importante debido a que los métodos utilizados para realizar los estudios de la Gaussianidad de procesos, se basan en la propiedad de estacionariedad de los procesos.

Se ha realizado este estudio[7] de las componentes (x, y, z) que forman cada una de las estaciones indicadas en la tabla anterior (Cuadro 1) obteniéndose los resultados del Cuadro 2:

Estación	**Box Test**	**ADF Test**	**KPSS Test**
Todas las estaciones	$< 2.2e\text{-}16$	< 0.01	> 0.01

Cuadro 2: Resultados de los test de independencia y estacionariedad

Observando la tabla, vemos que en cada una de las estaciones, las series temporales cumplen los requisitos necesarios para realizar el estudio, es decir, por los p-valores obtenidos rechazamos la hipótesis de independencia con el Box-Pierce and Ljung-Box test, rechazamos la hipótesis de no estacionariedad con el Augmented Dickey-Fuller test y no tenemos evidencias suficientes para rechazar la hipótesis nula de estacionariedad con el test restante. Por lo tanto, tenemos que los datos recogidos en cada estación son estacionarios y dependientes.

5.2 Resultados de los Test de Normalidad

En la tabla siguiente, resumimos los resultados obtenidos al realizar los test de Epps y Lobato y Velasco[8] para determinar si los procesos estocásticos siguen un modelo Gaussiano sin haber realizado la proyección del conjunto de datos:

[7] realizado en los ficheros adjuntos de nombres CoordX.R, CoordY.R y CoordZ.R

[8] Estas pruebas se han realizado utilizando el fichero SinProyectar.m

Estación	Epps Test	Lobato-Velasco Test
028N	0.6439	0.7845
028W	0.5815	0.4840
028V	0.8714	0.2660
029N	0.3736	0.2110
029W	0.9720	0.4686
029V	0.1459	0.1971
036N	0.4616	0.4695
036W	0.3382	0.3679
036V	0.1801	0.5836
094N	0.1060	0.0529
094W	**0.0128**	**0.0073**
094V	0.9742	0.5428
181N	0.7631	0.1122
181W	**0.0497**	**1.2061e-07**
181V	0.4974	0.6922
222N	0.2067	0.6181
222W	**0.0037**	**0.0011**
222V	0.0750	0.2737
224N	0.4777	0.6528
224W	0.7286	0.0946
224V	0.4660	**6.8434e-04**
225N	0.9587	0.6235
225W	0.8826	0.9797
225V	0.7156	0.9639
226N	0.5203	**0.0212**
226W	0.2524	**6.3209e-11**
226V	0.8788	**0.0113**
234N	0.7640	0.0730
234W	**0.0444**	0.0594
234V	0.6403	0.2593
433N	**0.0152**	0.1794
433W	0.4614	0.5157
433V	0.8618	0.9559

Cuadro 3: P-valores de los test de Epps y Lobato-Velasco sin proyectar en cada una de las estaciones

Analizando la tabla, observamos que nos encontramos cuatro casuísticas distintas:

1. **P-valor $<$ 0.05 en el test de Epps y en el test de Lobato y Velasco**.
   Notemos que para los procesos estocásticos 94W, 181W y 222W se ha obtenido un *P-valor $<$ 0.05* para los dos test. Luego, es condición suficiente para rechazar la hipótesis nula de Gaussianidad a nivel α.

   [1.a] P-valor $<$ 0.05 en el test de Epps. Para los procesos estocásticos 234W y 433N se ha obtenido un *P-valor $<$ 0.05* para el test de Epps. Como se ha obtenido al menos en uno de los test de Gaussianidad que el *P-valor $<$ 0.05*, esto es condición suficiente para rechazar la hipótesis nula de Gaussianidad a nivel α.

   [1.b] P-valor $<$ 0.05 en el test de Lobato y Velasco.
   Para los procesos estocásticos 224V, 226N, 226V y 226W se ha obtenido un *P-valor $<$ 0.05* para el test de Lobato y Velasco. Como se ha obtenido al menos en uno de los test de Gaussianidad que el *P-valor $<$ 0.05*, esto es condición suficiente para rechazar la hipótesis nula de Gaussianidad a nivel α.

2. **P-valor > 0.05 en el test de Epps y en el test de Lobato y Velasco**.
   En los 25 procesos estocásticos restantes se ha obtenido que en los dos tests el p-valor es mayor que 0.05, por lo que no tenemos evidencias suficientes para rechazar la hipótesis nula de Gaussianidad.

Entonces, según los resultados obtenidos al realizar este procedimiento, ya podemos obtener una primera conclusión y es que, las las series temporales 094W , 181W, 222W, 224V, 226N, 226W, 226V, 234W y 433N rechazan la hipótesis nula de Gaussianidad y, por tanto, no siguen una distribución normal (vease Cuadro 3). Para dar consistencia a estos resultados, vamos a realizar el FDR [13]. Con respecto a los procesos pertenecientes al punto 2, es posible que las variables sigan una distribución normal, mientras que el conjunto de variables multidimensionales no siga una distribución de este tipo. Por este motivo, se va a estudiar la normalidad de los procesos estocásticos utilizando los test anteriormente citados habiendo realizado previamente la proyección aleatoria de cada uno de los conjuntos de datos.

A continuación, se muestran los resultados de los test realizados sin el uso de la proyección aleatoria y el FDR utilizado para dar mayor consistencia a esos resultados:

Estación	**Epps Test**	**L-V Test**	**Hochberg**	**Yekutieli**
094W	**0.0128**	**0.0073**	0.0064	0.0192
181W	**0.0497**	**1.2061e-07**	2.4122e-07	3.6183e-07
222W	**0.0037**	**0.0011**	0.0019	0.0033
224V	0.4660	**6.8434e-04**	0.0014	0.0021
226N	0.5203	**0.0212**	0.0424	**0.0636**
226W	0.2524	**6.3209e-11**	1.2642e-10	1.8963e-10
226V	0.8788	**0.0113**	0.0226	0.0339
234W	**0.0444**	0.0594	0.0297	**0.0891**
433N	**0.0152**	0.1794	0.0304	0.0456

Cuadro 4: P-valores de los test de Epps y Lobato-Velasco sin proyectar en cada una de las estaciones y con FDR

Como se puede observar claramente en el *Cuadro 4*, los test de Epps y Lobato-Velasco sin proyección de datos para las estaciones mencionadas rechazan la hipótesis nula de Gaussianidad de manera consistente, ya que en al menos uno de los dos procedimientos del FDR el p-valor calculado es menor que 0.05. Resulta interesante observar que, para las estaciones 226N y 234W, el nuevo p-valor obtenido en Yekutieli es mayor que 0.05. Con estos resultados, concluimos que el método de Hochberg utilizando los datos sin proyectar es más consitente que el de Yekutieli.

A partir de ahora, nos centraremos en aquellos procesos para los cuales no hemos podido rechazar la hipótesis nula de Gaussianidad. Para estas series temporales, realizaremos el estudio de los test de Gaussianidad aplicados a los datos proyectados[9] seleccionando un vector aleatorio h con la distribución $\beta(A, B)$. Se ha realizado un análisis exhaustivo de los valores de A y B de la distribución β anterior para obtener resultados consistentes, sin embargo, únicamente se van a mostrar los p-valores correspondientes a los pares (A,B) para los que se han obtenido los resultados más significativos en cuanto al rechazo de la hipótesis nula. En las dos tablas siguientes se recopilan los resultados obtenidos para cada una de las estaciones:

[9]Las pruebas se han realizado utilizando el fichero Proyectando.m

Estación	**Test**	$\beta(2,7)$	$\beta(1,100)$	$\beta(100,1)$	$\beta(1,1000)$	$\beta(1,5000)$	$\beta(1,4500)$
028N	Epps	0.2150	0.0807	0.6179	0.1730	0.6395	0.0621
	L-V	0.4984	0.4569	0.7714	0.4659	0.7825	0.2988
028W	Epps	0.9793	0.8280	0.6652	0.6175	0.8934	0.9536
	L-V	0.8975	0.8517	0.4935	0.4289	0.9962	0.9802
028V	Epps	0.8661	0.9828	0.8287	0.5624	0.7599	0.6754
	L-V	0.1206	**0.0260**	0.2510	**0.0043**	0.0990	0.0536
		$\beta(50,150)$	$\beta(1,100)$	$\beta(2,7)$	$\beta(100,1)$	$\beta(1,1000)$	$\beta(1,200)$
029N	Epps	0.2766	0.2599	0.2251	0.3299	0.3150	0.2330
	L-V	0.1675	0.0989	0.0998	0.1750	0.1407	0.0886
029W	Epps	0.8965	0.7959	0.8783	0.9738	0.9420	0.8503
	L-V	0.3606	0.3723	0.3694	0.4532	0.4063	0.3641
029V	Epps	**0.0423**	**0.0497**	0.1512	0.1531	0.0980	0.1931
	L-V	0.1544	0.0664	**0.0490**	0.1935	**0.0294**	**0.0271**
		$\beta(100,1)$	$\beta(2,7)$	$\beta(1,100)$	$\beta(50,150)$	$\beta(100,200)$	$\beta(500,2000)$
036N	Epps	0.4616	0.9453	0.6882	0.9536	0.9464	0.9490
	L-V	0.4413	0.5012	0.3719	0.4510	0.4397	0.4582
036W	Epps	0.3136	0.3861	0.4666	0.3926	0.4039	0.4280
	L-V	0.3697	0.4881	0.3205	0.4635	0.4882	0.4559
036V	Epps	0.1919	0.4417	0.3754	0.4525	0.4828	0.4817
	L-V	0.6242	0.7311	0.4265	0.8339	0.8314	0.8339
		$\beta(50,150)$	$\beta(1,100)$	$\beta(2,7)$	$\beta(100,1)$	$\beta(1,200)$	$\beta(1,400)$
094N	Epps	0.1977	0.1714	0.2387	0.1203	0.2068	0.4263
	L-V	0.0931	**0.0471**	0.0904	0.0560	0.4039	0.3932
094V	Epps	0.5507	0.6673	0.4906	0.9554	0.7366	0.5030
	L-V	0.6478	0.6969	0.6980	0.5398	0.4761	0.7197
		$\beta(100,1)$	$\beta(2,7)$	$\beta(1,100)$	$\beta(1,200)$	$\beta(50,100)$	$\beta(1,1000)$
181N	Epps	0.7533	0.3557	0.3255	0.3082	0.3579	**0.0410**
	L-V	0.1452	0.2837	0.4702	0.45714	0.8277	**0.0006**
181V	Epps	0.4630	0.9539	0.7799	0.9363	0.8707	0.9678
	L-V	0.7275	0.9811	0.6213	0.9513	0.9873	0.9184
		$\beta(100,1)$	$\beta(2,7)$	$\beta(1,100)$	$\beta(2,100)$	$\beta(50,100)$	$\beta(1,1000)$
222N	Epps	0.2063	0.3989	0.6305	0.5028	0.3950	0.5898
	L-V	0.6114	0.5066	0.5076	0.4787	0.4947	0.3316
222V	Epps	0.1226	0.4582	0.3985	0.9481	0.5873	0.8765
	L-V	0.3186	0.7868	0.7266	0.9969	0.8276	0.7334
		$\beta(100,1)$	$\beta(2,7)$	$\beta(1,100)$	$\beta(1,200)$	$\beta(1,1000)$	$\beta(5,200)$
224N	Epps	0.4055	0.0752	**0.007**	0.0686	**0.0148**	**0.0368**
	L-V	0.6584	0.5924	0.1806	0.1472	0.1235	0.2021
224W	Epps	0.7326	0.6215	0.3626	0.5047	0.5396	0.6606
	L-V	0.0961	0.3977	0.2500	0.2857	0.3758	0.3429

Cuadro 5: P-valores obtenidos al realizar el test de Epps y Lobato y Velasco para los datos proyectados aleatoriamente de las estaciones 028, 029, 036, 094, 181, 222 y 224

		$\beta(100,1)$	$\beta(2,7)$	$\beta(1,100)$	$\beta(1,200)$
225N	Epps	0.8905	0.4494	0.3105	0.2996
	L-V	0.598	**0.0089**	**0.0004**	**0.0257**
225W	Epps	0.9021	0.1066	0.0672	0.0573
	L-V	0.9714	0.2503	0.2945	**0.0243**
225V	Epps	0.9476	0.4417	0.5746	0.3200
	L-V	0.6139	**0.0199**	**0.03063**	**0.0023**
		$\beta(100,1)$	$\beta(2,7)$	$\beta(1,100)$	
234N	Epps	0.7230	0.4179	0.0720	-
	L-V	0.1047	0.7067	0.1790	-
234V	Epps	0.5498	0.3452	0.5030	-
	L-V	0.2063	**0.0380**	**0.0262**	-
		$\beta(100,1)$	$\beta(2,7)$	$\beta(1,100)$	
433W	Epps	0.5096	0.8232	0.1688	-
	L-V	0.5095	0.6621	0.5459	-
433V	Epps	0.8624	0.5010	0.4289	-
	L-V	0.9254	0.7522	0.8885	-

Cuadro 6: P-valores obtenidos al realizar el test de Epps y Lobato y Velasco para los datos proyectados aleatoriamente de las estaciones 225, 234 y 433

En primer lugar, centrémonos en el *Cuadro 5*. Como se puede observar, para la estación de Santa Mónica bay 028 no se ha encontrado ningún par *(A,B)* para el que se haya podido rechazar la hipótesis nula de Gaussianidad para el test de Epps. Mientras que, para el test de Lobato y Velasco, encontramos posibles rechazos en la coordenada 028V con los valores de A y B siguientes: $(1,100)$ y $(1,1000)$.

En el mismo cuadro, podemos encontramos los resultados correspondientes a la estación de Point Reyes 029, observamos que existen posibles rechazos únicamente para la coordenada 029V. Por una parte, los valores de A y B para los que podríamos rechazar la hipótesis nula del test de Epps, son $(50,150)$ y $(1,100)$. Por otra parte, los valores correspondientes a los posibles rechazos con el test de Lobato y Velasco son $(2,7)$, $(1,1000)$ y $(1,200)$.

Para las series temporales que forman la estación de Cape Mendocino 094, también incluidos en el *Cuadro 5*, podemos identificar que únicamente se ha obtenido un posible valor de rechazo. En este caso, corresponde al test de Epps con el par $(1,100)$ en la coordenada 094N. Centrémonos ahora a la estación Rincón 181. Se puede identificar fácilmente que para el valor de A y B $(1,1000)$ se ha encontrado un p-valor que permitiría rechazar la hipótesis nula tanto del test de Epps como la de Lobato y Velasco. Si observamos ahora la estación 224 - Wallops Island, se identifica que únicamente obtenemos posibles rechazos del test de Epps para la coordenada 224N con los valores $(1,100)$ y $(1,1000)$ y $(5,200)$. Ahora bien, fijándonos en el *Cuadro 6*, se ve claramente que sucede lo contrario con las series temporales de la estación de Kaneohe Bat 225. Se puede detectar que únicamente se podría rechazar la hipótesis de Gaussianidad para el test de Lobato y Velasco en las tres coordenadas. Para las coordenadas 225N y 225V, rechazaríamos con los pares *(A,B)* siguientes: $(2,7)$, $(1,100)$ y $(1,200)$. Mientras que, para la coordenada 225W, rechazaríamos únicamente con el valor de $(A,B) = (1,200)$. En la estación Santa Bárbara 234 incluida también en el *Cuadro 6*, encontramos que es probable el rechazo de la gaussianidad del test de Lobato y Velasco para la serie temporal correspondiente a la coordenada 234V y los parámetros (A,B) : $(2,7)$ y $(1,100)$.

Por último, si nos fijamos en las estaciones: Grays Harbor 036 (*Cuadro 5)*, Lucía Escarpment 222 (*Cuadro 5*) y Duck FRF 433 (*Cuadro 6*), podemos observar que no ha sido posible identificar, ni para el test de Epps ni para el test de Lobato y Velasco, ningún valor significativo en cuanto a posibles rechazos de la hipótesis nula.

Nótese que se ha hecho mucho hincapié en que cuando el p-valor obtenido al realizar este procedimiento para los dos test, Epps y Lobato y Velasco, se ha tratado como *posible* valor significativo ya que, cuándo realizamos la proyección aleatoria a un nivel de significación del 95 %, va a exisitir siempre la posibilidad de que el p-valor obtenido al ejecutar el test esté dentro del 5 % que rechaza la hipótesis nula erróneamente y, por tanto, sería un falso positivo.

Es por ello por lo que, para poder concluir corretamente sobre si los datos que estamos estudiando siguen o no una distribución Gaussiana, se ha realizado el estudio de los test de Epps y Lobato y Velasco múltiple para nuestros datos proyectados y, dado que siempre existe un error residual, se ha aplicado el test de ajuste o FDR de Benjamini - Hochberh y el de Benjamini-Yekutili detallados en la sección 6.

En las siguientes tablas se presentan los resultados obtenidos, donde el campo *Mayor i de rechazo* hace referencia al indicador del mayor p-valor para el cual se puede rechazar la hipótesis de Gaussianidad:

Estación	Test	FDR	Nuevo p_0	Rechazo	Mayor i de rechazo
028N	Epps	Hochberg	0.00056	Sí	1000
		Yekutieli	1.85427	No	-
	L-V	Hochberg	0.00094	Sí	1000
		Yekutieli	4.90900	No	-
028W	Epps	Hochberg	0.00075	Sí	1000
		Yekutieli	5.55261	No	-
	L-V	Hochberg	0.00051	Sí	1000
		Yekutieli	3.79824	No	-
028V	Epps	Hochberg	0.00100	Sí	1000
		Yekutieli	7.47932	No	-
	L-V	Hochberg	0.00076	Sí	1000
		Yekutieli	1.40198	No	-
029N	Epps	Hochberg	0.00068	Sí	1000
		Yekutieli	2.60979	No	-
	L-V	Hochberg	0.00030	Sí	1000
		Yekutieli	1.22587	No	-
029W	Epps	Hochberg	0.00100	Sí	1000
		Yekutieli	7.45544	No	-
	L-V	Hochberg	0.00065	Sí	1000
		Yekutieli	3.51016	No	-
029V	Epps	Hochberg	0.00049	Sí	1000
		Yekutieli	1.55481	No	-
	L-V	Hochberg	0.00061	Sí	1000
		Yekutieli	0.65432	No	-
036N	Epps	Hochberg	0.00096	Sí	1000
		Yekutieli	7.20526	No	-
	L-V	Hochberg	0.00046	Sí	1000
		Yekutieli	3.42076	No	-
036W	Epps	Hochberg	0.00089	Sí	1000
		Yekutieli	5.24871	No	-
	L-V	Hochberg	0.00077	Sí	1000
		Yekutieli	3.95487	No	-
036V	Epps	Hochberg	0.00024	Sí	1000
		Yekutieli	1.58510	No	-
	L-V	Hochberg	0.00069	Sí	1000
		Yekutieli	4.97996	No	-
094N	Epps	Hochberg	0.00066	Sí	1000
		Yekutieli	1.92797	No	-
	L-V	Hochberg	0.00037	Sí	1000
		Yekutieli	1.20832	No	-
094V	Epps	Hochberg	0.00097	Sí	1000
		Yekutieli	7.29060	No	-
	L-V	Hochberg	0.00054	Sí	1000
		Yekutieli	4.06253	No	-

Cuadro 7: Resultados al realizar el ajuste del test multiple de Epps y Lobato y Velasco para los datos proyectados aleatoriamente de cada una de las estaciones 028, 029, 036 y 094

Estación	Test	FDR	Nuevo p_0	Rechazo	Mayor i de rechazo
181N	Epps	Hochberg	0.00097	Sí	1000
		Yekutieli	1.99144	No	-
	L-V	Hochberg	0.00100	Sí	1000
		Yekutieli	0.29934	No	-
181V	Epps	Hochberg	0.00050	Sí	1000
		Yekutieli	3.71807	No	-
	L-V	Hochberg	0.00081	Sí	1000
		Yekutieli	6.02530	No	-
222N	Epps	Hochberg	0.00022	Sí	1000
		Yekutieli	1.62230	No	-
	L-V	Hochberg	0.00062	Sí	1000
		Yekutieli	4.61777	No	-
222V	Epps	Hochberg	0.00016	Sí	1000
		Yekutieli	0.97335	No	-
	L-V	Hochberg	0.00031	Sí	1000
		Yekutieli	2.29141	No	-
224N	Epps	Hochberg	0.00099	Sí	1000
		Yekutieli	0.69356	No	-
	L-V	Hochberg	0.00098	Sí	1000
		Yekutieli	4.79213	No	-
224W	Epps	Hochberg	0.00100	Sí	1000
		Yekutieli	5.64956	No	-
	L-V	Hochberg	0.00004	Sí	1000
		Yekutieli	0.00030	Sí	1
225N	Epps	Hochberg	0.00090	Sí	1000
		Yekutieli	5.36116	No	-
	L-V	Hochberg	0.00001	Sí	1000
		Yekutieli	0.00009	Sí	247
225W	Epps	Hochberg	0.00066	Sí	1000
		Yekutieli	1.31571	No	-
	L-V	Hochberg	0.00086	Sí	1000
		Yekutieli	0.22848	No	-
225V	Epps	Hochberg	0.00100	Sí	1000
		Yekutieli	5.92155	No	-
	L-V	Hochberg	0.00099	Sí	1000
		Yekutieli	7.09407	No	-

Cuadro 8: Resultados al realizar el ajuste del test multiple de Epps y Lobato y Velasco para los datos proyectados aleatoriamente de cada una de las estaciones 181, 222, 224 y 225

Estación	Test	FDR	Nuevo p_0	Rechazo	Mayor i de rechazo
234N	Epps	Hochberg	0.00088	Sí	1000
		Yekutieli	4.84025	No	-
	L-V	Hochberg	0.00088	Sí	1000
		Yekutieli	5.70780	No	-
234V	Epps	Hochberg	0.00098	Sí	1000
		Yekutieli	4.45763	No	-
	L-V	Hochberg	0.00039	Sí	1000
		Yekutieli	0.30093	No	-
433W	Epps	Hochberg	0.00100	Sí	1000
		Yekutieli	7.48004	No	-
	L-V	Hochberg	0.00099	Sí	1000
		Yekutieli	5.02693	No	-
433V	Epps	Hochberg	0.00100	Sí	1000
		Yekutieli	6.19492	No	-
	L-V	Hochberg	0.00100	Sí	1000
		Yekutieli	7.46140	No	-

Cuadro 9: Resultados al realizar el ajuste del test multiple de Epps y Lobato y Velasco para los datos proyectados aleatoriamente de cada una de las estaciones 234 y 433

Si se observa el *Cuadro 7*, el *Cuadro 8* y el *Cuadro 9*, es fácil comprobar que al realizar el Test múltiple el ajuste de Benjamini-Hochberg es mucho más consistente que el de Yekutieli, ya que este último únicamente rechaza la hipótesis de Gaussianidad de las series temporales 224W y 225N para el test de Lobato y Velasco.

6 Conclusiones

Las evidencias que se han probado anteriormente, demuestran que los resultados de los test de Epps y Lobato y Velasco son importantes cuando se tiene series estacionarias y dependientes. En nuestro estudio, hemos considerado una muestra de los procesos estocásticos que satisficieran dichas condiciones. Para encontrar de manera más sencilla aquellos procesos que no siguen una distribuación Gaussiana, se ha realizado tanto el test de Epps como el de Lobato y Velasco de las series temporales obteniendo los siguientes resultados:

- Para las series temporales 094W, 181W, 222W, 224V, 226N, 226W, 226V, 234W y 433N, se ha conseguido rechazar la hipótesis nula de Gaussianidad, ya que el p-valor resultante de realizar el test de Epps o el test de Lobato y Velasco a resultado menor que 0.05. Esto implica, que en un primer estudio se ha conseguido obtener que un 27 % de los procesos estocásticos no siguen un distribución normal.

- Por otra parte, no se han obtenido evidencias suficientes para rechazar la hipótesis nula de Gaussianidad para las 24 series temporales restantes. Por ello, se ha procedido a realizar un estudio más exhaustivo sobre estos conjuntos de datos.

Para realizar el estudio más detallado del que se ha hablado en el último punto, se ha utilizado la proyección aleatoria aplicada al test múltiple. Somos conscientes de que los métodos para ajustar la tasa de falsos positivos de Hochberg y Yekutili han sido creados para comparar hipótesis distintas. Sin embargo, resulta interesante aplicar estos métodos a una misma hipótesis, pues ofrece la ventaja de que, al realizar la prueba, si al menos una de las múltiples hipótesis se rechaza, se podrá concluir que el proceso es no Gaussiano. Dado este punto, se han sacado las siguientes conclusiones:

- En la selección de nuestra distribución $\beta(A, B)$ hemos observado que los parámetros A y B no siguen ningún patrón, lo cuál nos ha llamado la atención ya que esperábamos que al aumentar o disminuir los valores tuviera una cierta relación con en el hallazgo de un p-valor menor que 0.05.

- Aplicando el método de Hochberg al test de Epps y al de Lobato y Velasco, hemos conseguido obtener el rechazo de las hipótesis nulas en cada una de las coordenadas que mide nuestra boya.
- El caso opuesto ocurre con el método de Yekutieli, únicamente se han podido rechazar las hipótesis de Gaussianidad para las estaciones 224W y 247N.

Consecuentemente, como con el método de Hochberg se rechazan todas las hipótesis nulas, se puede concluir que la altura de las olas del mar no sigue una distribución Gaussiana. Además, podemos decir que el método de Hochberg es mucho más consistente que Yekutieli. Este método sirve para cualquier tipo de dependencia positiva y no nos rechaza siempre la hipótesis nula de Gaussianidad. No obstante, Hochberg no es aplicable para cualquier tipo de dependencia. Sin embargo, entendemos que cumplimos con las condiciones para poder confiar en este método.

Para complementar y enriquecer a la vez este trabajo se propone:

- Estudiar que las series temporales correspondientes a la altura de las olas del mar satisfacen las condiciones de dependencia de Hochberg.
- Realizar un análisis análogo tomando el conjunto de datos en $\mathbb{R}^3$.

Referencias

[1] Abramson, J. School of Mathematical and Statistical Sciencies. Arizona State University. Recuperado el 15 de Mayo de 2019 de https://math.libretexts.org/

[2] Azencott, R and Dacunha-Castelle, D. (1986). Series of Irregular Observations: Forecasting and Model Building. *Springer.*

[3] Benjamini, Y. and Hochberg, F. 1995, Journal of the Royal Statistical Society. Series B (Methodological), 57 (1),289-300.

[4] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**(4), 1165-1188.

[5] Bingham, E. and Mannila, H. (1999). Random projection in dimensionality reduction.

[6] Box, G.E.P. y Pierce D.A.(1970). Distribution of Residual Autocorrelations in Autoregressive-Integrated Movin Average Time Series Models. *Journal of the American Statistical Association.***65**(332), 1509-1526.

[7] Croarkin, C. y Tobias, P.(2012). Engeniering statistics.

[8] Cuesta Albertos, J.A. Cálculo de Probabilidades.

[9] Cuesta-Albertos,J.A., Del Barrio,T., Fraiman, R. y Matrán, C. (2007). The random projection method in goodness of fit for functional data. *Comput. Statist. Data Anal.* **51**(10), 4814-4831.

[10] Dickey, D y Fuller, W. (1979).Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association.***74**(366). 427-431.

[11] Dr. Richard, J., Seymour,David Castel, Dr. Robert Guza and Associate PI Dr. William O'Reilly. Monitoring and Prediction of Waves and Shoreline Change. *CDIP: Coastal data information program.* Recuperado el 15 de Mayo de 2019 de https://cdip.ucsd.edu/

[12] Epps, T. W. (1987). Testing that a stationary time series is Gaussian. *Ann. Statist.* **15**(4), 1683-1698.

[13] Friedman, J., Hastie, T.m Tibshirani, R. (2008). The elements of statistical learninf. *Data Mining,Inference and Prediction Ann. Statist.* **15**(4), 1683-1698.

[14] Hossein Pishro, N. Introduction to Probability, statistics and Random Processes. *Joint Distributions.* Recuperado el 16 de Mayo de 2020 de https://www.probabilitycourse.com

[15] Kwiatkowski D. et al. (1991). Testing the null hypothesis of trend stationarity

[16] Lobato, I.N y Velasco,C.(2004). A simple test of normality for time series. *Econometric Theory.***20**(4), 671-689

[17] Molina, I. Series Temporales. *Procesos estocásticos Estacionarios*

[18] Newcastle University. Academic Skills Kit. Recuperado el 20 de Mayo de 2020 de https://internal.ncl.ac.uk/

[19] Nieto-Reyes, A., Cuesta-Albertos, J.A. y Gamboa, F. (2014). A random-projection based test of Gaussianity forstationary processes. 124-141.

[20] Nieto-Reyes, A. (2010). Aplicaciones Estadísticas de las Proyecciones Aleatorias.

[21] Paya, R. Variable compleja. *Funciones analíticas.*

[22] Sarabia Alegría, J.M. y Pascual Sáez, M. Curso básico de estadística para los grados en economía y administración y dirección de empresas. 185 - 208.

[23] Yekutieli, D. and Benjamini, Y. (1999). A resampling based false discovery rate controlling mul- tiple test procedure. *J Statist. Plann. Inference* **82**, 171-196.

A Anexo

Se incluye el código utilizado en R para estudiar la estacionariedad de los conjuntos de datos que tenemos

```
#Marta Ferrero Díez
install.packages('xlsx',dependencies = T)
# install.packages('rJava', dependencies = T)
# install.packages('xlsxjars', dependencies = T)
install.packages ('openxlsx')
library('readxl')
# install.packages("tseries")
install.packages('xlsx')
library(readxl)
library(xlsx)
library (openxlsx)
require (openxlsx)



# %%%%%%ESTACION POINT REYES 029 %%%%%%
#Cargamos nuestros datos
d1=read.xlsx(file = "E:\\TFG\\TFG_Alicia\\TFG_Alicia\\e_029.xlsx",sheetIndex = 1)
e_029N=d1[,2]

#Dibujamos la gráfica de los datos de la coordenada y
plot(1:length(e_029N),e_029N,'line')
hist(e_029N)
hist(rnorm(length(e_029N),mean(e_029N),sd(e_029N)))

## Pasamos ahora los siguientes tests:
# Box.test
# adf.test
# kpps.test
# Lobato-Velasco, Epps (con Matlab)

# Aplicamos  Box.test , adf.test y kpss.test para ver si la muestra de los datos e
# es estacionaria y dependiente. Estacionario significa que cualquier vector tiene
# distribución cuando se traslada en el tiempo.

# Pasamos el Box.test: Box-Pierce tests que habla de independencia :
#  H0: son independientes
#  Ha: son dependientes
#nosotros lo que queremos es rechazar la hipótesis nula.

#INDEPENDENCIA

Box.test(e_029N)
#p-value < 2.2e-16 <0.05 luego rechazamos la hipótesis nula , luego no son indepen

#ESTACIONALIDAD

#Ahora que sabemos que tenemos dependencia miramos la estacionariedad, pues querem
```

```
#con datos dependientes y estacionarios
#Para ello pasamos el test adf y el kpss:

# adf.test : Augmented Dickey-Fuller test que nos muestra la estacionariedad :
# H0: no estacionario
# Ha: estacionario
#primero cargamos el paquete tseries
#install.packages('tseries', dependencies = T)
library('tseries')

adf.test(e_029N)
#p-value< 0.01 <0.05 rechazamos la hipotesis nula de no estacionariedad
#alternative hypothesis: stationary


#Pasamos otro test de estacionariedad. kpss.test.

# test kpss.test : Kwiatkowski-Phillips-Schmidt-Shin test
# H0: proceso estacionario
# Ha: No es estacionario
kpss.test(e_029N, null = 'Trend')
#p-value >0.1 > 0.05 no tenemos evidencias suficientes para rechazar la hipótesis

kpss.test(e_029N, null = 'Level')
#p-value>0.1 > 0.05 no tenemos evidencias suficientes para rechazar la hipótesis n
#Luego con todos los test pasados podemos concluir que nuestro conjunto de datos e
```

Se utiliza el siguiente script para la creación del Cuadro 3. En el script se llama a la función test.m que a su vez llama a las funciones RealDataG.m y RealDataE.m

```
%PARTE NO PROYECTADA EPPS Y LV
clear all
clc
format short e

%Inicializamos los datos

e_028= xlsread('e_028.xlsx');
e_029= xlsread('e_029.xlsx');
e_036= xlsread('e_036.xlsx');
e_094= xlsread('e_094.xlsx');
e_181= xlsread('e_181.xls');
e_222= xlsread('e_222.xls');
e_224= xlsread('e_224.xlsx');
e_225= xlsread('e_225.xlsx');
e_226= xlsread('e_226.xls');
e_234= xlsread('e_234.xls');
e_433= xlsread('e_433.xlsx');

datos28N=e_028(:,2)';
datos28W=e_028(:,3)';
datos28V=e_028(:,4)';
datos29N=e_029(:,2)';
datos29W=e_029(:,3)';
datos29V=e_029(:,4)';
datos36N=e_036(:,2)';
datos36W=e_036(:,3)';
datos36V=e_036(:,4)';
datos94N=e_094(:,2)';
datos94W=e_094(:,3)';
datos94V=e_094(:,4)';
datos181N=e_181(:,2)';
datos181W=e_181(:,3)';
datos181V=e_181(:,4)';
datos222N=e_222(:,2)';
datos222W=e_222(:,3)';
```

```
datos222V=e_222(:,4)';
datos224N=e_224(:,2)';
datos224W=e_224(:,3)';
datos224V=e_224(:,4)';
datos225N=e_225(:,2)';
datos225W=e_225(:,3)';
datos225V=e_225(:,4)';
datos226N=e_226(:,2)';
datos226W=e_226(:,3)';
datos226V=e_226(:,4)';
datos234N=e_234(:,2)';
datos234W=e_234(:,3)';
datos234V=e_234(:,4)';
datos433N=e_433(:,2)';
datos433W=e_433(:,3)';
datos433V=e_433(:,4)';

%Vamos a empezar pasando los test Lobato y Velasco y Epps sin proyección,
%esto es, estudiamos si las variables por separado son normales, el test
%consiste en las siguientes hipótesis :
%H0: las variables Xi son normales
%Ha: las variables Xi no son normales
%Sin embargo puede ocurrir que las variables por separado
%X1,X2,...,Xn sean normales pero en conjunto (X1,X2,...,Xn) no lo sean
%por lo tanto , aplicaremos después los tests proyectando, es decir
%estudiamos si (X1,X2,...,Xn) sigue una distribución normal.
%Proyectamos tomando distintos valores para A y B en ConProyeccion
 v=[datos28N;datos28W;datos28V;datos29N;datos29W;datos29V;datos36N;datos36W;datos36V;datos94N;
     datos94W;datos94V;datos181N;datos181W;datos181V;datos222N;datos222W;datos222V;
     datos224N;datos224W;datos224V;datos225N;datos225W;datos225V;datos226N;datos226W;datos226V
         ;datos234N;datos234W;datos234V;datos433N;datos433W;datos433V]';
[filv,colv]=size(v);


%== SIN PROYECTAR ==

ve=[];
vl=[];
for i=1:colv
[ TE(i),PvalueE(i),TLv(i),PvalueLv(i) ] = test(v(:,i));
ve=[ve;PvalueE(i)];
vl=[vl;PvalueLv(i)];
end
pvaloresTotales=[ve vl]
[filpv,colpv]=size(pvaloresTotales);
hrechazadasE=[];
hrechazadasLV=[];
for j=1:filpv
    if pvaloresTotales(j,1)<0.05
        hrechazadasE=[hrechazadasE;pvaloresTotales(j,1)];
        phrechazadasE=length(hrechazadasE)/filpv*100;
        fprintf('Un %2f porciento de las H0 de Gaussianidad son rechazadas con el Test Epps\n'
            ,phrechazadasE)
    end
     if pvaloresTotales(j,2)<0.05
        hrechazadasLV=[hrechazadasLV;pvaloresTotales(j,2)]
        phrechazadasLV=length(hrechazadasLV)/filpv*100;
        fprintf('Un %2f porciento de las H0 de Gaussianidad son rechazadas con el Test LV\n',
            phrechazadasLV)
    end

end



%%CREAMOS TABLA DE PVALORES SIN PROYECTAR
t=uitable;
cnames={'Test de Epps','Test de LV'};
rnames={'28N','28W','28V','29N','29W','29V','36N','36W','36V','94N','94W','94V','181N','181W',
    '181V','222N','222W','222V',...
    '224N','224W','224V','225N','225W','225V','226N','226W','226V','234N','234W','234V','433N'
        ,'433W','433V'};
 set(t,'Data',pvaloresTotales,'ColumnName',cnames,'RowName',rnames)
```

```
function [ TE,PvalueE,TLv,PvalueLv ] = test( x )

```

```
n=length(x);

[TE, PvalueE]=RealDataE(x)
if PvalueE >0.05
    disp('No hay evidencias suficientes para rechazar la H0 de Epps : Las variables siguen una
        distribución normal')
else
    disp('Rechazamos H0 : Las variables no siguen una distribución normal')
end
[TLv , PvalueLv]=RealDataG(x)

if PvalueLv >0.05
    disp(' No hay evidencias suficientes para rechazar la H0 de LV: Las variables siguen una
        distribución normal')
else
    disp('Rechazamos H0 : Las variables no siguen una distribución normal')

end
```

```
function [T , Pvalue]=RealDataG(x)
T=GestadisticoVn(x,1); Pvalue=(1-chi2cdf(T,2));
```

```
% Input data:  x = is the process we want to test, it is given in a
% row vector
% Output data: Pvalue = pvalue obtained by doing the random
% projection test to the process x
function [T , Pvalue]=RealDataE(x)
n=length(x); N=2; dN=2*N; rn=floor(n^.4);
dev=std(x)*(n-1)/n;
T=Sub([1  2]/dev,x,dev,rn,n,dN,N);
Pvalue=1-chi2cdf(T,2);
```

Las funciones GestadisticoVn.m y Sub.m han sido obtenidos de [20].

Para la realización del Cuadro 4 hasta el Cuadro 16, se ha utilizado el siguiente fichero .m que a su vez llama a Hochberg2 y Yekutili2.m:

```
%PARTE PROYECTANDO CON EPPS Y LV y FDR
%Parte de Matlab.
clear all
clc
format short e

% Inicializamos los datos

datos= xlsread('x');

%Con R hemos visto que era estacionario.
long=length(datos);

%== PROYECTADOS ==
% Aplicamos ahora los datos proyectados :
disp('Datos proyectados')
%Se realiza el tést de hipótesis múltiple con n=1000
n=1000;

vector1LVP=zeros(n,1);
vector1EP=zeros(n,1);
cont=0;
contE=0;

[TLV,TE,PvalorLVP,PvalorEP]=RealDataLByEP(datos,A,B)

for i=1:n
    [TLB,TE,PvalorLVP,PvalorEP]=RealDataLByEP(datos,A,B);
    vector1LVP(i)=PvalorLVP;
    vector1EP(i)=PvalorEP;
end
disp('FALSE DISCOVERY RATE')

%%%%%REALIZAMOS EL FDR CON HOCHBERG)%%%%%%%

m=length(vector1EP);
r=length(vector1LVP);
```

```
alpha=0.05;
fprintf('====== BENJAMINI-HOCHBERG ====== \n')
[ ordenadoH1, pcorregidos1, ih1, p0h1] = Hochberg2(vector1EP',alpha ) ;
fprintf('El p-valor del proceso de Benjamini Hochberg y  Epps es %5.5f  \n',p0h1)
if ih1>0
fprintf('Con Hochberg y el test de EPPs de un total de 1000 hipótesis se rechazan las H(i) con
      de %5g hasta %5g \n',1, ih1)
else
    disp('No se rechaza ninguna hipótesis')
end

[ ordenadoH2, pcorregidos2,ih2 ,p0h2] = Hochberg2(vector1LVP',alpha );
fprintf('El p-valor del proceso de Benjamini Hochberg y  LV es %5.5f  \n',p0h2)

if ih2>0
fprintf('Con Hochberg y el test de LV de un total de 1000 hipótesis se rechazan las H(i) con i
     de %5g hasta %5g \n',1,ih2)
else
    disp('No se rechaza ninguna hipótesis')
end
%%%%REALIZAMOS EL FDR CON YEKUTIELI%%%%%%%

fprintf('====== BENJAMINI-YEKUTIELI ====== \n')
[ ordenadoY1, pcorregidosY1,iy1,p0y1 ] = Yekutili2(vector1EP',alpha);
fprintf('El p-valor del proceso de Benjamini Yekutieli y  Epps es %5.5f  \n',p0y1)
if iy1>0
fprintf('Entonces rechazamos las hipótesis nulas H(i) desde %5g hasta %5g \n',1, iy1)
else
    disp('No se rechaza ninguna hipótesis')
end

[ ordenadoY2, pcorregidosY2,iy2,p0y2 ] = Yekutili2(vector1LVP',alpha);
fprintf('El p-valor del proceso de Benjamini Yekutieli y  LV es %5.5f  \n',p0y2)
if iy2>0
fprintf('Entonces rechazamos las hipótesis nulas H(i) desde %5g hasta %5g \n',1,iy2)
else
    disp('No se rechaza ninguna hipótesis')
end
```

```
function [ ordenado, pcorregidos,i,p0 ] = Hochberg2(x,alpha )
%Se aplica el método de Hochberg para realizar el ajuste del testing
%múltiple
%   x= vector fila
m=length(x);
ordenado=sort(x);


nv=ordenado./[1:m];
constante=(m+1-[1:m]);
operacion=constante.*nv;
minimo=min(operacion);
p0=minimo;



for i=1:m
    pcorregidos(i)=(i/(m+1-i))*alpha;
end
i=m;
 while ordenado(i)>=pcorregidos(i)
     i=i-1;
   end

end
```

```
function [ ordenado, pcorregidos,i,p0 ] = Yekutili2(x,alpha )
%UNTITLED4 Summary of this function goes here
%   x= vector fila
m=length(x);
ordenado=sort(x);


nv=ordenado./[1:m];
minimo=min(nv);
parte1=m*(sum(1./[1:m]));
```

```
p0=parte1*minimo;


for i=1:m
    pcorregidos(i)=i*(alpha)/parte1;
end
i=m;
 while i>0 && ordenado(i)>=pcorregidos(i)
     i=i-1;
   end
% if i==0

end
```